Site-Restricted Web Searches for Data Collection in Regional Dialectology

Jack Grieve, Centre for Forensic Linguistics, Aston University, Birmingham, UK
Costanza Asnaghi, English Linguistics, Università Cattolica del Sacro Cuore, Milan, Italy
Tom Ruette, QLVL, University of Leuven, Belgium

Submitted to *American Speech*, July 3rd, 2012

Corresponding Author:
Dr. Jack Grieve
Lecturer in Forensic Linguistics
School of Languages and Social Sciences
Aston University, Birmingham, UK
j.grieve1@aston.ac.uk

Acknowledgments

We would like to thank Douglas Biber, Patricia Cukor-Avila**,** Aaron Dinkin, Dirk Geeraerts, Inoue Fumio**,** Wilbert Heeringa, Bill Kretzschmar, Danny Long, John Nerbonne, Dirk Speelman, Joeri Theelen, Emily Waibel, and Martijn Wieling for their comments on this paper and the methods presented in this paper and a preliminary version of this paper, which was presented at *Methods in Dialectology 14* at the University of Western Ontario, August 3rd, 2011. We would also especially like to thank Bert Vaux for making the data from the Harvard Dialect Survey available to us for comparison.

Site-Restricted Web Searches for Data Collection in Regional Dialectology

Abstract

This paper presents a new method for data collection in regional dialectology that is based on site-restricted web searches. This method allows for the values of many lexical alternation variables to be measured across a region using common search engines such as Google or Bing. The method involves estimating the proportions of the variants of a lexical alternation variable over a series of cities by counting the number of webpages that contain these variants on newspaper websites originating from these cities through site-restricted web searches. This method is evaluated by mapping nine content word alternations with known distributions in American English. In almost all cases, the maps based on the data gathered through site-restricted web searches align closely with traditional dialect maps based on data gathered through linguistic interviews, demonstrating that this method allows for regional lexical variation to be measured accurately. Unlike collecting dialect data through linguistic interviews, however, which can take years to complete, the use of site-restricted web searches allows for dialect data to be collected from across the United States in a matter of days.

1. Introduction

Regional dialect studies are usually based on data gathered through linguistic interviews. The linguistic interview can take several forms, but in general it involves the linguist initiating a communicative event to elicit language data from a particular informant. In dialectology, most surveys have gathered data by eliciting specific linguistic forms based on a questionnaire (Kurath, 1949), although running speech has also been sampled (Szmrecsanyi, 2008), and occasionally a combination of these two techniques for data collection have been used (Labov et al, 2006). Questionnaires are usually administered in person (Kurath, 1949), but in American dialect surveys questionnaires have also been administered by mail (Davis, 1948), over the phone (Labov et al, 2006), and online (Vaux, 2003). Although interviewing individual informants is a valid method for collecting dialect data, conducting interviews with informants across a sufficiently large and geographically dispersed set of locations is a very time-consuming process. As such data collection for all major American dialect surveys has taken years to complete and has usually only been based on the language of a few informants at each location.

This paper presents a new method for data collection in regional dialectology that is based on site-restricted web searches. This method can be used to quickly measure regional variation in the values of many lexical alternation variables using a common search engine such as Google or Bing. Basically, to measure the proportion of the variants of an alternation variable over a series of cities, the number of webpages in which these variants appear are counted on newspaper websites originating from these cities through a series of site-restricted web searches. For example, the alternation between *sneakers* and *tennis shoes* could be measured in Andalusia, Alabama by counting the number of webpages on the *andalusiastarnews.com* website that contain each of these words based on the results of two site-restricted web searches made on Google: *site:andalusiastarnews.com "sneakers"* and *site:andalusiastarnews.com "tennis shoes"*. The proportion of *sneakers* in Andalusian newspaper writing would then be calculated by dividing the number of hits for *sneakers* by the number of hits for *sneakers* plus the number of hits for *tennis shoes.* This process would then be repeated for newspapers in many cities from across the United States to identify patterns of regional linguistic variation in *sneakers/tennis shoes* alternation. The goal of this paper is to introduce and evaluate this method for the

collection of regional lexical variation data by mapping nine lexical alternation variables with clear and established patterns of regional variation in American English.

This paper is organized as follows. First, the harvesting of 1,349 American newspaper websites is described. Based on this list of newspaper websites, the proportion of *sneakers*, *tennis shoes, running shoes* and *gym shoes* was measured in cities from across the United States through site-restricted web searches. The analysis of this one alternation variable is presented in detail in order to exemplify the application of the method. The use of the spatial autocorrelation statistic local Getis-Ord *Gi* is then introduced through an analysis of this data. This statistic identifies significant patterns of spatial clustering in the values of a variable measured over a series of locations, allowing for underlying regional patterns to be identified in the data collected through site-restricted web searches. Finally, the method is evaluated by mapping nine lexical alternations variables with known distributions in American English based on the results of the *Harvard Dialect Survey* (Vaux, 2003). This comparison shows that this method for the collection of dialect data allows for linguistic variables to be accurately mapped in a fraction of the time needed to collect data through linguistic interviews.

2. Newspaper Selection

The basic method for the data collection being introduced here involves estimating the values of a linguistic variable across a series of locations based on web searches that are restricted to websites originating from these locations. In this paper, the method is evaluated by mapping content word alternations with known distributions in American English through web searches that are restricted to websites for newspapers from across the United States.

In order to access American newspaper websites, a list of over 2,000 newspapers was harvested from the website *refdesk.com*, along with the city, state and URL associated with each of these newspapers. This particular newspaper index was selected because it was well organized and simply designed, which facilitated data harvesting, and because it appeared to list a relatively large number of newspaper websites compared to similar websites. After the websites were harvested, the *www.* prefix was stripped from each URL to allow for additional URLs associated with the newspapers to be

accessed through site-restricted web searches (e.g. allowing for *topics.nytimes.com* to be searched in addition to *www.nytimes.com).* Each of the URLs was then tested online and approximately half of the URLs were discarded because they were inactive or because they were not associated with a sizeable number of webpages. In addition, a small number of business, entertainment, and university newspapers were deleted from the list in order to focus the analysis on the typical local newspaper register of American English. The list was then checked by hand to see if the largest cities and most popular newspapers in the United States were represented. If a city or newspaper was missing, a newspaper URL was manually added to the list whenever possible. In addition, the cities represented by the newspapers were mapped and regional gaps were filled by adding newspapers from the largest cities in those regions whenever possible. In total, the final version of the list contains 1,349 newspaper websites representing 1,232 cities from across the contiguous United States.

3. The Measurement of *Sneakers/Tennis Shoes/Running Shoes/Gym Shoes* Alternation

This section introduces the use of site-restricted web searches for data collection in regional dialectology through a detailed analysis of the alternation between the synonyms *sneakers, tennis shoes, running shoes* and *gym shoes.* This lexical alternation variable is suitable for analysis because its variants are relatively frequent and generally interchangeable in newspaper writing. An alternation variable that does not meet these criteria is not suitable for measurement using this basic method (although see Section 6). These four variants were selected for analysis because they are the most frequent variants for this concept in American English according to the Harvard Dialect Survey (see Section 5.1). The decision to exclude less frequent variants, including *trainers, runners* and *jogging shoes,* is justified below. In addition to being suitable for analysis using site-restricted web searches, the analysis of this content word alternation was selected to exemplify the application of the method because it allows for both the analysis of multi-word lexical items and the analysis of an alternation consisting of more than two variants to be discussed.

At the core of the method being introduced here is the use of site-restricted web searches. When querying Google (or Bing), a search can be restricted to websites whose URLs contain a particular

string by including that string prefixed with the *site:* tag in the search box in addition to the search string. For example, searching for *site:nytimes.com "tennis shoes"* counts the number of webpages on the *nytimes.com* website that contain "tennis shoes."[1] The basic idea behind this method for data collection is to use site-restricted web searches to count the number of webpages in which the variants of an alternation variable occur in hundreds of newspaper websites from across a region of interest. The search engine can be queried manually but it is much easier to query the search engine automatically using computer programs designed for harvesting information online. In this case, a Perl LWP script was written to automatically download the html source code from the URL associated with the results page for that web search (e.g. *http://www.google.com/search?&q=%22tennis+shoes%22&sitesearch=nytimes.com*) and the number of hits were then extracted from this html code. It is necessary, however, to limit the speed at which Google is queried because Google will block an IP address if it submits too many queries. The basic approach adopted here was to query Google approximately 200 times in row with a random interval between searches of between 1 and 10 seconds and then to take a 20 minute break. Although the searches were made for this analysis using Google, it is much easier and quicker to use Bing, which uses simpler html code and places fewer restrictions on the frequency of searches. Based on informal comparisons, it appears to make very little difference whether Google or Bing is used.

Before conducting a full analysis, however, it is important to ensure that the site-restricted web searches are primarily identifying interchangeable uses of the variants under analysis. This can be achieved by looking over some of the webpages listed on the results pages generated by the site-restricted web searches. For example, nine of the first ten webpages found by searching for *sneakers* on *nytimes.com* linked to newspaper articles where *sneakers* could have been replaced with the other variants, including an article on sneakers that tone your leg muscles while you walk and an article on a pair of sneakers designed to commemorate the World Basketball Festival. The other webpage, however, contained information on the 1992 movie "Sneakers." This hit is problematic because *sneakers* cannot be replaced with the other variants in this context because *sneakers* is being used as a proper noun. If the most common variants of an alternation variable are highly polysemous or commonly used as a part

of proper nouns or idioms, then the variable is probably not suitable for analysis using site-restricted web searches, at least using the basic method being introduced here (although see Section 6). In this case, however, non-interchangeable uses were relatively infrequent for all four of the variants.

The frequency, however, of non-interchangeable uses of *runners* and *trainers* made it necessary to exclude these variants from the analysis. The most common uses of these terms are not synonymous with *sneakers* and *tennis shoes*, with *runners* most often referring to people who run and with *trainers* most often referring to people who train. It is acceptable to exclude these variants from the analysis because these variants, as well as other variants such as *jogging shoes* and *athletic shoes*, are rarely used in American English to refer to *sneakers* according to the *Harvard Dialect Survey*, where these variants account for less that 1% of the total response to this item on the questionnaire. The decision to exclude these variants from the analysis does violate the principle of accountability, which requires that all variants be considered when measuring an alternation variable (Labov, 1972). However, from a statistical standpoint, as long as the most common variants are considered, the principle of accountability does not need to be adhered to because the exclusion of low frequency variants cannot have any substantial effect on the proportion of high frequency variants. For example, if only the two most frequent variants were analyzed here, by calculating the proportions of *sneakers* and *tennis shoes* and excluding *running shoes* and *gym shoes*, the resultant maps would have been almost identical to the maps for *sneakers* and *tennis shoes* presented above, despite the fact that *running shoes* and *gym shoes* accounted for over 20% of the total hits. Excluding even less frequent variants is therefore entirely acceptable.

The number of hits for *sneakers*, *tennis shoes*, *running shoes* and *gym shoes* was therefore measured across the 1,349 newspaper websites through a series of automated site-restricted web searches on Google. Overall, *sneakers* was found to be the most common variant accounting for 54% of the total hits, *tennis shoes* was the second most common variant accounting for 25% of the total hits, *running shoes* was the third most frequent variant accounting for 13% of the total hits, and *gym shoes* was the least frequent variant, accounting for 8% of the total hits.

The data was then pooled for every city that was represented by two or more newspapers. This

was accomplished by summing the hit counts for each variant for all the newspapers from the same city. The proportion of each variant was then calculated for the 880 cities for which at least one hit for at least one variant was counted by dividing the number of hits for that variant by the total number of hits for all variants for that city, in this case yielding four sets of proportions measured over 880 cities[2]. It was necessary to exclude the other 352 cities from this analysis because none of the variants occurred in these newspapers. It is important to note that there is a general lack of agreement in dialectology and sociolinguistics about how to measure non-binary alternation variables, especially when measured quantitatively and especially when the variants cannot be arranged in a natural ranking, as is generally the case for lexical alternation variables (see Chambers and Trudgill, 1998). Measuring the proportion of each variant separately relative to all variants is a simple solution to this problem.

The proportions of each variant were then mapped across the 880 cities based on the longitude and latitude for each city as defined by the US Postal Service[3]. These maps are presented in Figures 1-4. The lightness of a dot represents the proportion of the variant at that location: a lighter dot indicates that the variant is relatively common at that location and a darker dot indicates that the variant is relatively uncommon at that location. Figure 1 shows that *sneakers* is most common in the Northeast, whereas Figure 2 shows that *tennis shoes* is most common across the rest of the United States, especially in the Southeast and the North. Figures 3 and 4 are less clear, with Figure 3 appearing to show that *running shoes* is most common in the West, and with Figure 4 appearing to show that *gym shoes* is most common in the Midwest. There is, however, no need to rely on a subjective analysis to determine if these variables are regionally patterned; the statistical analysis presented in Section 4 will allow for these preliminary observations to be verified by identifying the locations of significant high- and low-value clusters for each set of proportions.

Before describing this statistical analysis, it is important to consider why the maps do not exhibit clearer regional patterns. It is undeniable that gathering dialect data using web searches that are restricted to newspaper websites will result in data that is much noisier than if data had been gathered through linguistic interviews. There are several reasons why this is the case. First, site-restricted web searches will always result in some non-interchangeable uses of the variants being counted, such as

webpages that reference the movie *Sneakers*. Second, it is difficult to fully control for register variation. Ideally, only webpages representing a specific register would be searched, such as newspaper articles, but this is not usually possible. For example, although the majority of the webpages with newspaper URLs contain newspaper articles, it is clear that many additional registers are also represented, including pages with comments and online information that are not found in a print newspaper. Furthermore, the range and proportion of registers found on different newspaper websites is bound to vary. Third, the era represented by each newspaper will also vary due to different chronological depths of newspaper archives. Fourth, web searches only allow for the number of webpages that contain a particular form to be counted, as opposed to the number of forms. Fifth, the hit counts returned by a search engine are only estimates which are unstable because search engines are constantly updated and newspapers webpages are regularly modified. Finally, site-restricted web searches do not usually allow for the demographic background of informants to be controlled. Most notably, given syndication practices, it is certain that a sizeable percentage of the webpages associated with a particular newspaper will not even be written by residents of the city or even the state where that newspaper is published.

All of these factors essentially introduce noise into the dataset, making it harder to find patterns of regional linguistic variation. The existence of these many sources of noise, however, does not necessarily invalidate the method. No approach to data collection is perfect. All approaches to data collection are affected by noise. The goal of this paper is to determine if this method for data collection is useful despite these problems, which will partially be overcome through the use of spatial autocorrelation statistics, as described in Section 4.

Figure 1 Proportion of *Sneakers* (Web Searches)

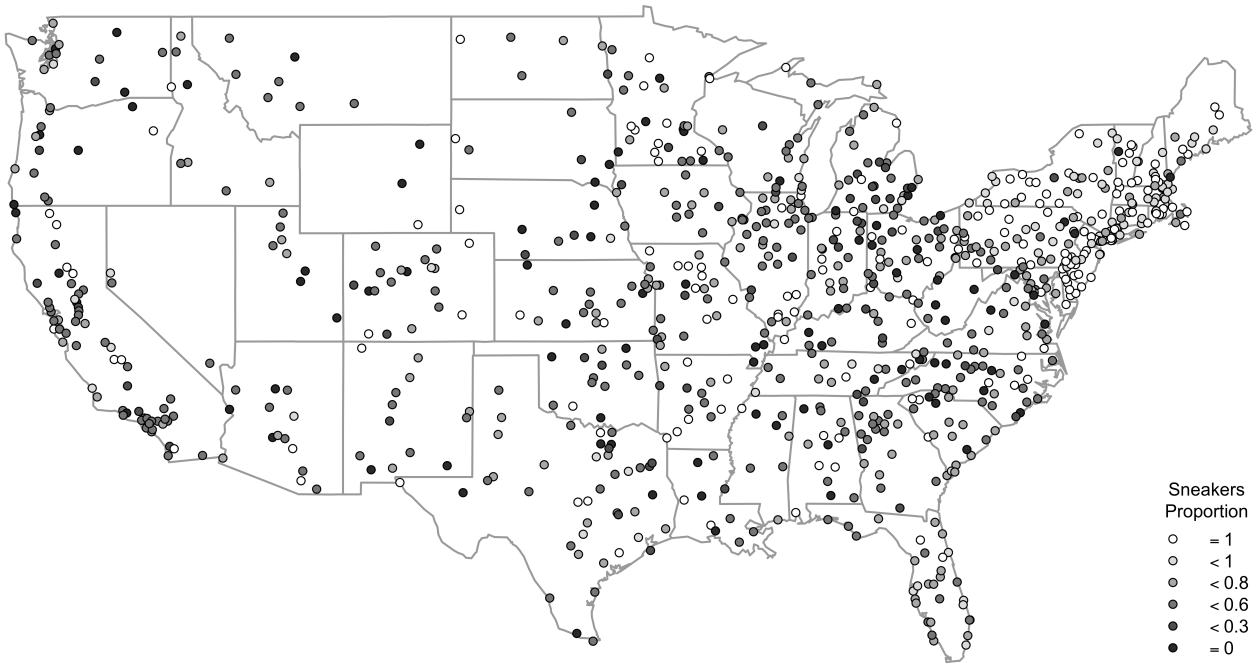


Figure 2 Proportion of *Tennis Shoes* (Web Searches)

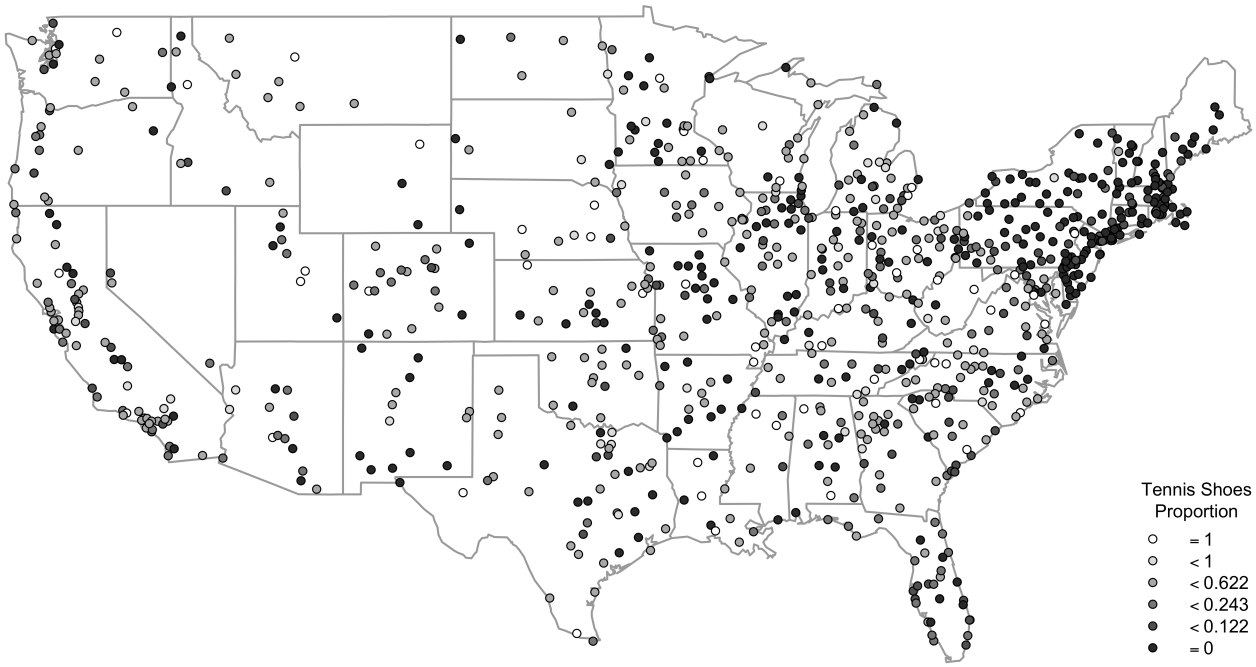


Figure 3 Proportion of *Running Shoes* (Web Searches)

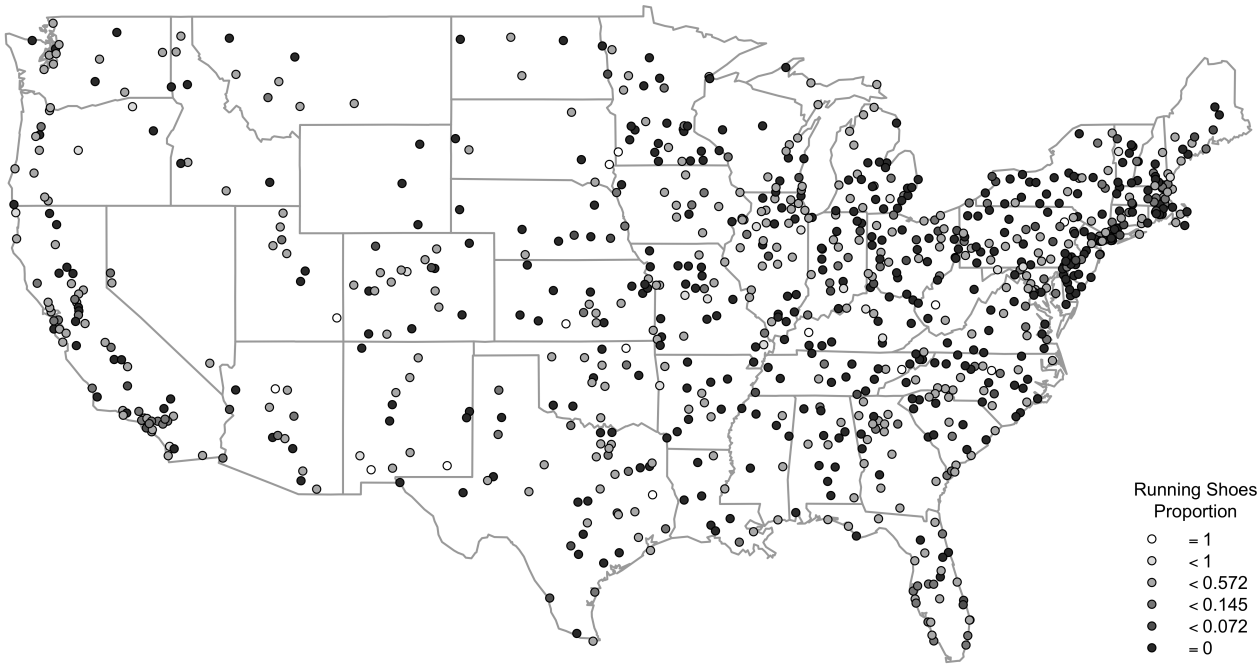
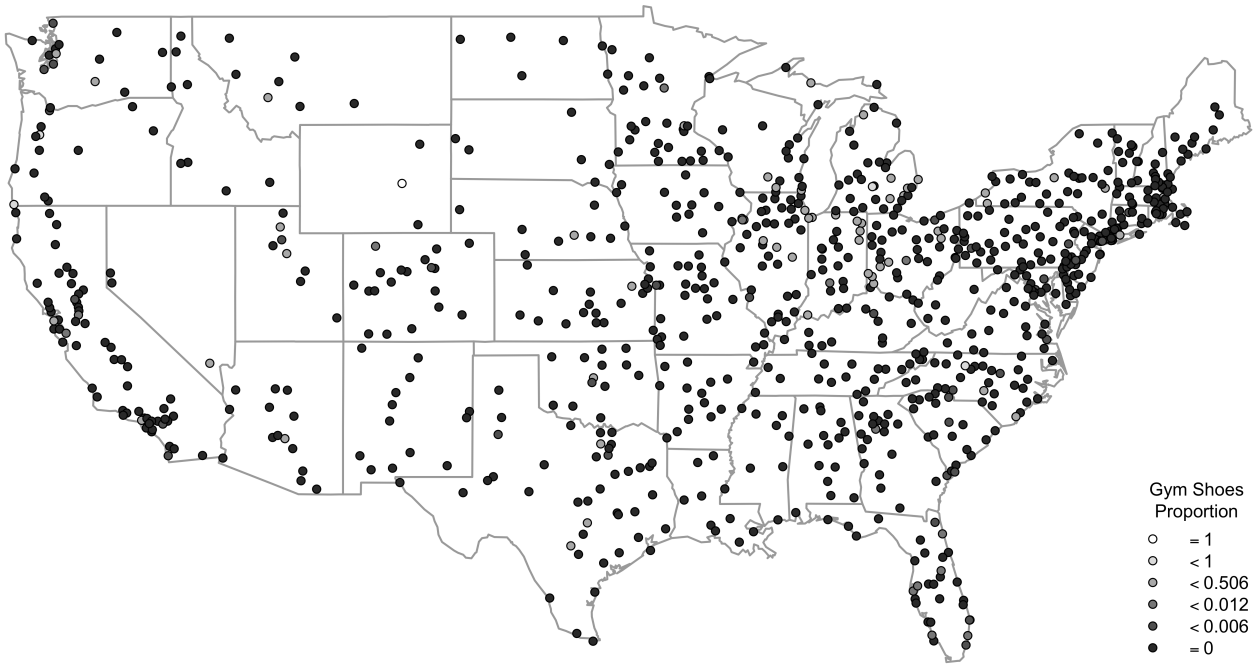


Figure 4 Proportion of *Gym Shoes* (Web Searches)



4. Spatial Autocorrelation Analysis

Because the raw dialect maps do not exhibit clear patterns of regional variation, a local spatial autocorrelation analysis (Grieve, 2011; Grieve et al, 2011) was used to identify significant patterns of spatial clustering for each variant.[4] In particular, each variant was subjected to a local Getis-Ord *Gi* analysis (Ord and Getis, 1995) to identify clusters of locations where each variant is especially common. By comparing the proportion of each variant at each location to the proportions of that variant at nearby locations, a Getis-Ord *Gi* analysis produces a *z*-score for each location indicating the degree to which that location is part of a high value cluster (significant positive z-score), a low-value cluster (significant negative z-score), or a region of variability (a non-significant z-score approaching zero). The results of the local spatial autocorrelation analysis are then mapped to identify the locations of these clusters thereby allowing for underlying patterns of regional variation to be identified while controlling for the various sources of non-regional noise introduced through data collection. This is essentially a statistical method for plotting isoglosses.

To calculate Getis-Ord *Gi*, it is necessary to define a *spatial weighting function*, which is a set of rules that assigns a weight to the comparison of every pair of locations where comparisons between locations that are close together are given greater weight than comparisons between locations that are far apart (Odland, 1988). The analyses reported here are based on a reciprocal weighting function (see Grieve, 2011), which is a common weighting function that assigns a weight to a comparison based on the reciprocal of the distance between the two locations, so that weighting decreases with distance (Odland, 1988). In addition, given the large number of locations in the dataset, each location was only compared to the closest 300 locations. A range of other spatial weighting functions were tested but in general varying the spatial weighting function had very little effect on the results of the analysis.

The local autocorrelation maps for these four variables are plotted in Figures 5-8. In these maps, clusters of lighter circles represent regions where the variant under analysis is most common—not necessarily compared to the other variants but compared to the other locations for that variant—and clusters of darker circles represent regions where the variant under analysis is least common. It is entirely possible, however, that the variant is relatively infrequent in that region compared to other

more common variants. Note that it is possible for the clusters for different variants to overlap because in the regions where the infrequent variants are most common, the frequent variants may still be more common. Each of these maps focuses on locations with positive Getis-Ord Gi z-scores, as these are the locations where the variant under analysis is most common. The individual maps do not reveal which alternative variants are most common in the rest of the country; to determine which alternative variants are most common it is necessary to inspect the maps for these variants. When mapping binary alternations, a slightly different approach is used because in these cases it is possible to represent the clusters for both variants on the same map (see Section 5.2.2).

All of these maps identify clear and significant patterns of regional variation and confirm the subjective analysis of the raw maps presented above. Figure 5 shows that *sneakers* is most common in the Northeast, as locations in the northeast tend to have significant positive Getis-Ord *Gi z*-scores. Similarly, Figure 6 shows that *tennis shoes* is most common in the rest of the United States, especially in the Southeast and the North. Figure 7 shows a clear pattern for *running shoes*, which is most common in the Mountain States and the Great Plains. Finally, Figure 8 shows that *gym shoes* is most common in the Midwest and to a lesser extent in the Pacific Northwest, especially Oregon. The utility of the spatial autocorrelation analysis is demonstrated by the maps for these final two variants, where regional patterns were not as clear in the raw maps as they were for the first two variants. The local spatial autocorrelation analysis therefore allowed for significant underlying patterns of regional variation to be identified objectively—patterns that may have been overlooked in a traditional analysis.

Figure 5 Local Autocorrelation Map for *Sneakers* (Web Searches)

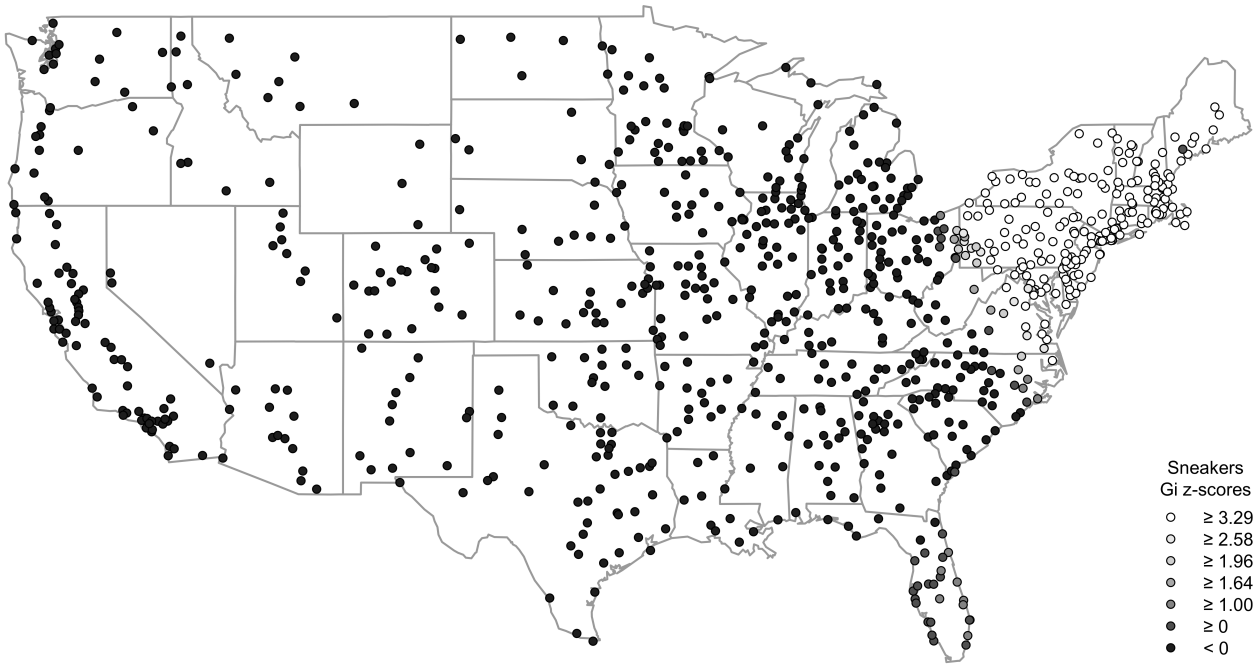


Figure 6 Local Autocorrelation Map for *Tennis Shoes* (Web Searches)

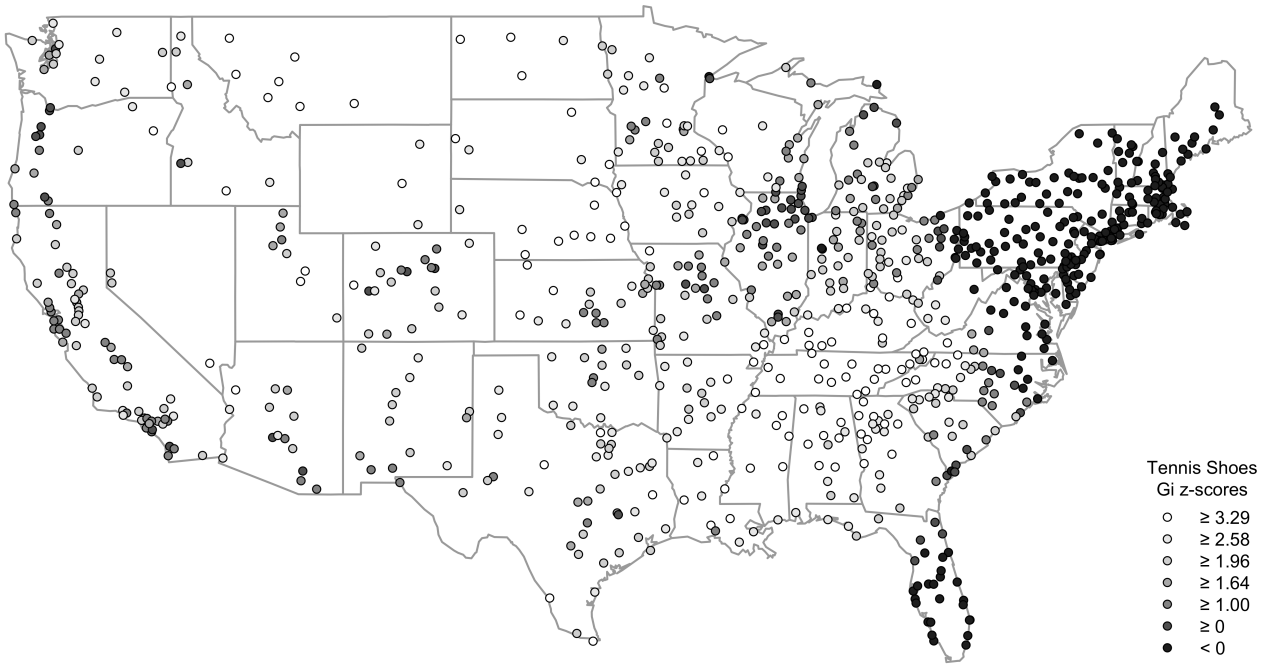


Figure 7 Local Autocorrelation Map for *Running Shoes* (Web Searches)

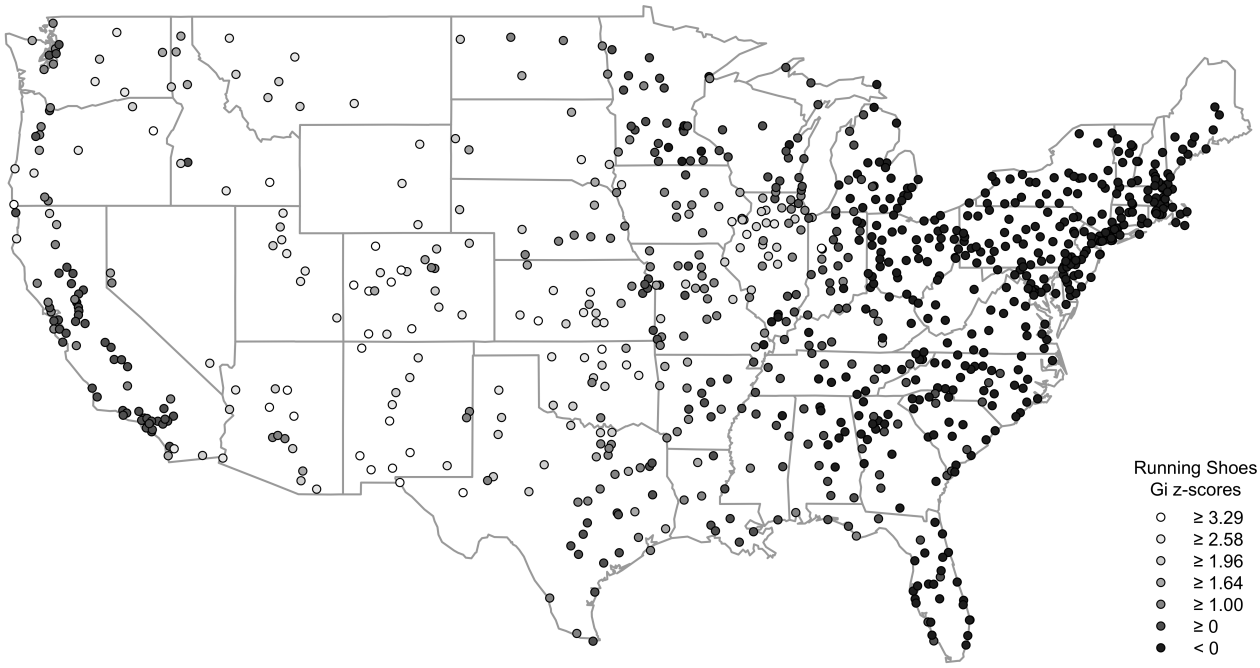
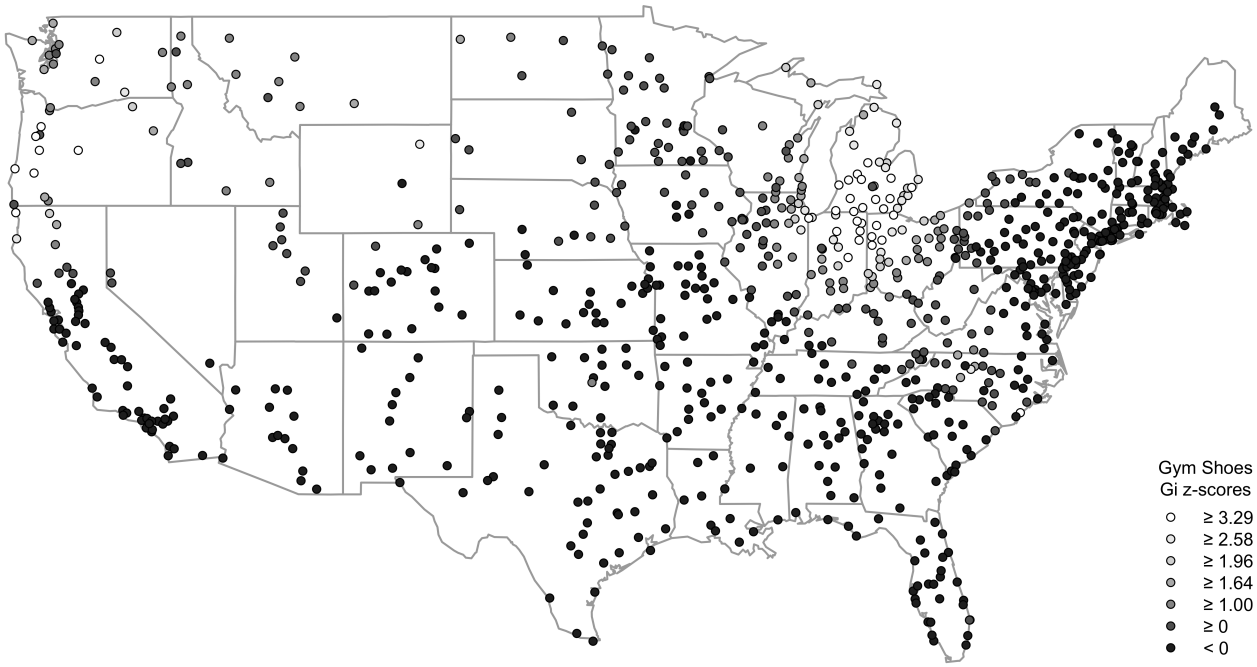


Figure 8 Local Autocorrelation Map for *Gym Shoes* (Web Searches)



5. Evaluation

To test this method for the observation of regional lexical variation, nine lexical alternation variables with known distributions in American English were measured across the 1,349 newspaper URLs and the results were then subjected to a spatial autocorrelation analysis and mapped. These maps were then compared to the results of the Harvard Dialect Survey (Vaux, 2003) to evaluate the method. Before presenting the results, however, the selection of these nine lexical alternation variables is discussed.

5.1 Variable Selection

Lexical alternation variables, specifically content word alternations, were selected for analysis based on two criteria. Most important, the variable must have been mapped in a previous American dialect survey. Second, the variable must be suitable for analysis using site-restricted web searches. In particular, the variables must consist of a set of variants that are synonymous in the majority of contexts and at least one variant must be relatively common in American newspaper writing.

There are not many content word alternations, however, that meet these two criteria, primarily because there have been only a few dialect surveys that have mapped content word alternations in American English. Furthermore, the two most important surveys of lexical variation in American English are not well suited for evaluating this method. The largest source of lexical data for the eastern United States is Hans Kurath's *A Word Geography of the Eastern United Sates* (1949). This data, however, is quite dated and comparable data is only available for New England (Kurath et al, 1939), the Upper Midwest (Allen, 1973), the Gulf Coast (Pederson, 1984-1992), and Texas (Atwood, 1962). The rest of the United States was never mapped and the data from these surveys were never combined. Furthermore, many of the alternations included in these surveys, such as farming terms, do not occur in newspaper writing. It is therefore impossible to use these datasets to verify the maps being generated here for the contiguous United States. The second major lexical dialect survey is the *Dictionary of American Regional English* (Cassidy & Hall, 1985, 1991; Hall & Cassidy, 1996; Hall, 2002; 2012; Carver, 1987). The *Dictionary*, however, does not provide proper maps for individual lexical items nor does it make its data publicly available. This survey therefore cannot be used to verify the method

either.

It was therefore necessary to select variables for analysis based on the *Harvard Dialect Survey* (HDS; Vaux, 2003)—the only survey that has mapped numerous everyday content word alternations in modern American English across the United States. The HDS was based on an online questionnaire that elicited 122 phonological, grammatical and lexical alternation variables, completed by over 47,000 informants between 2002 and 2003. The maps for all 122 items are available online[5]; however, there is no formal publication describing the methods and the results of HDS (although see Vaux, 2003). Nonetheless, the raw data from this survey were made available for analysis here by Bert Vaux. Although the data for the HDS is categorical in the sense that each informant is associated with a single variant for each item on the questionnaire, because in most cases there were many informants for each city, this dataset was quantified by calculating the proportion of informants from each city who preferred each variant of each lexical variable (for all cities with at least 5 informants). These proportions were then subjected to a local spatial autocorrelation analysis and mapped (as described above) so that the results of the HDS could be compared to the results obtained here.[6]

In particular, the nine lexical alternation variables from the HDS selected to evaluate the method are *sneakers/tennis shoes/running shoes/gym shoes*, *frosting/icing*, *trash can/garbage can*, *water fountain/drinking fountain*, *bag/sack*, *take-out/carry-out*, *cut the grass/mow the lawn/mow the grass*, *garage sale/yard sale/rummage sale/tag sale*, *grandmother/grandma/granny/nana*.

5.2 Results

The results presented here were gathered through site-restricted web searches made automatically on Google between October 14th and November 2nd, 2011 and then again between December 14th and December 29th, 2011. For each variable, a local spatial autocorrelation map was generated for each of its variants. These maps are presented here and then compared to the corresponding local autocorrelation maps based on the data from the HDS.

5.2.1 Sneakers/Tennis Shoes/Running Shoes/Gym Shoes

The first variable analyzed here is the alternation between *sneakers, tennis shoes, running shoes* and

gym shoes. The local spatial autocorrelation maps for each of these four variants were presented in Figures 5-8. These maps showed that *sneakers* is most common in the Northeast, *tennis shoes* is most common in most of the rest of the country except Florida, California and Illinois, *running shoes* is most common in the Central and Mountain States, and *gym shoes* is most common in the Midwest.

This analysis is largely confirmed by the HDS data. The local autocorrelation maps for the proportions of HDS informants who prefer each of the four variants in 1,162 cities are presented in Figures 9-12. The HDS maps for the two most common variants align closely with the maps based on the data gathered through site-restricted web searches. The HDS map for *running shoes*, however, identifies a larger *running shoes* region that encompasses all of the Central and Mountain States. This is perhaps because the HDS sampled fewer locations in this region and because the *running shoes* variant was much less common in the HDS data overall. The HDS map for *gym shoes* also aligns with the map based on the data gathered through the site-restricted web searches, although the *gym shoes* region is considerably larger and stronger in the HDS map.

Figure 9 Local Autocorrelation Map for *Sneakers* (Harvard Dialect Survey)

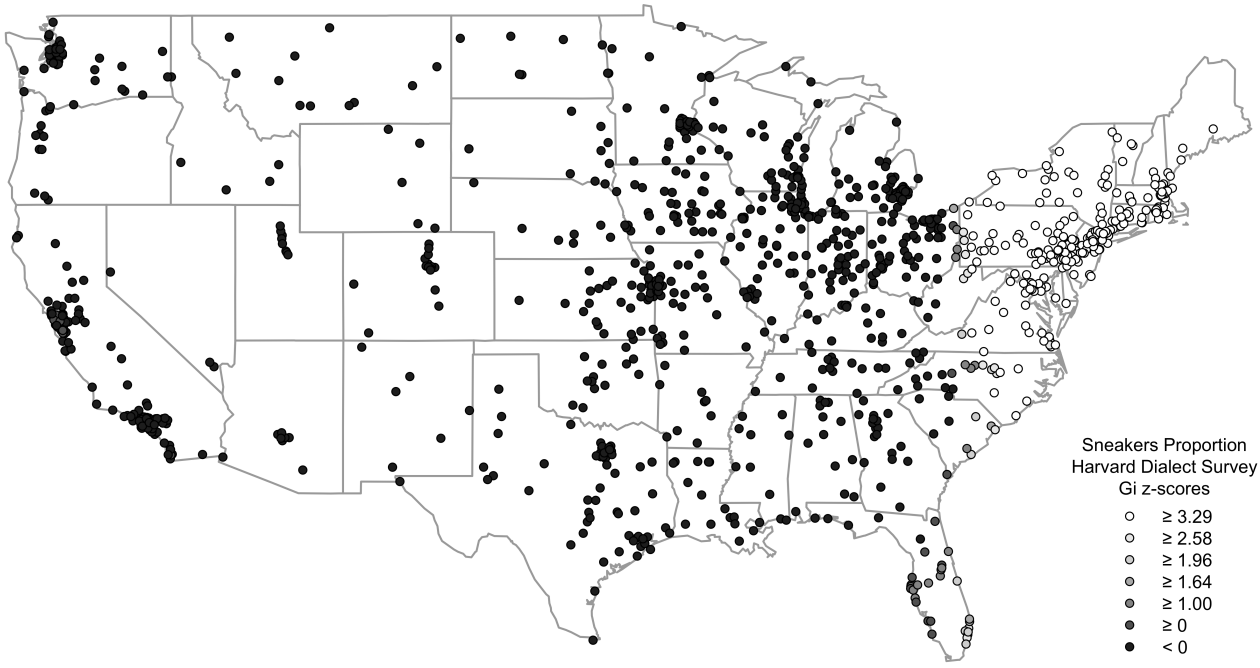


Figure 10 Local Autocorrelation Map for *Tennis Shoes* (Harvard Dialect Survey)

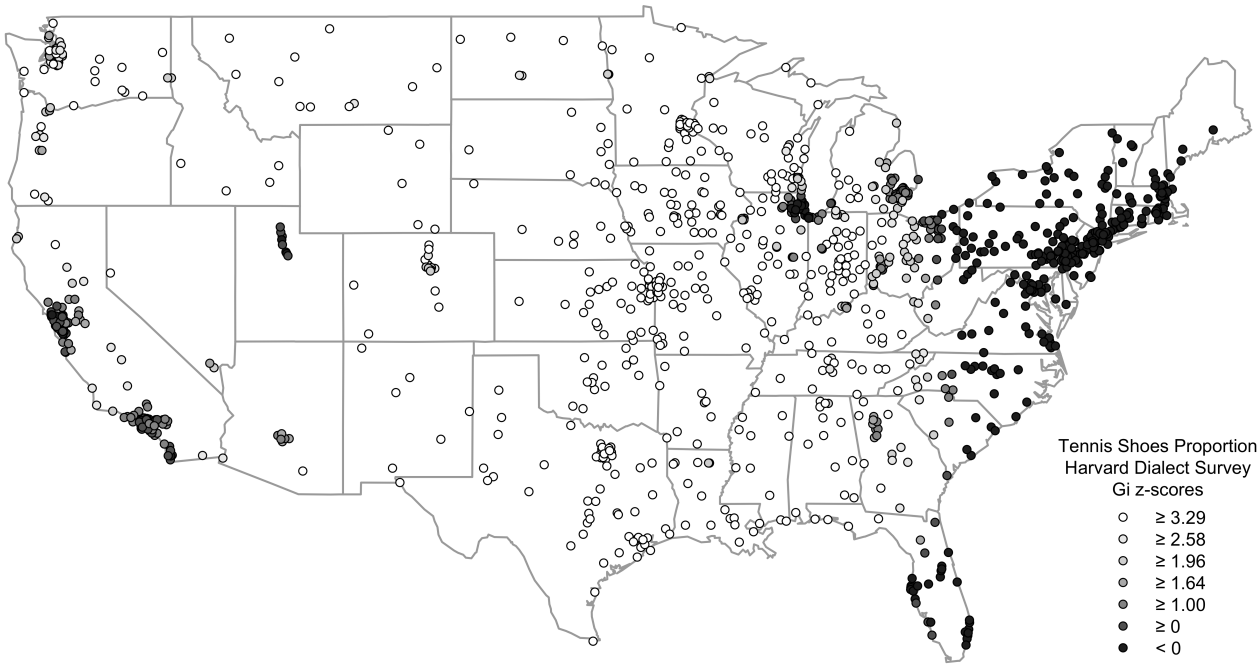


Figure 11 Local Autocorrelation Map for *Running Shoes* (Harvard Dialect Survey)

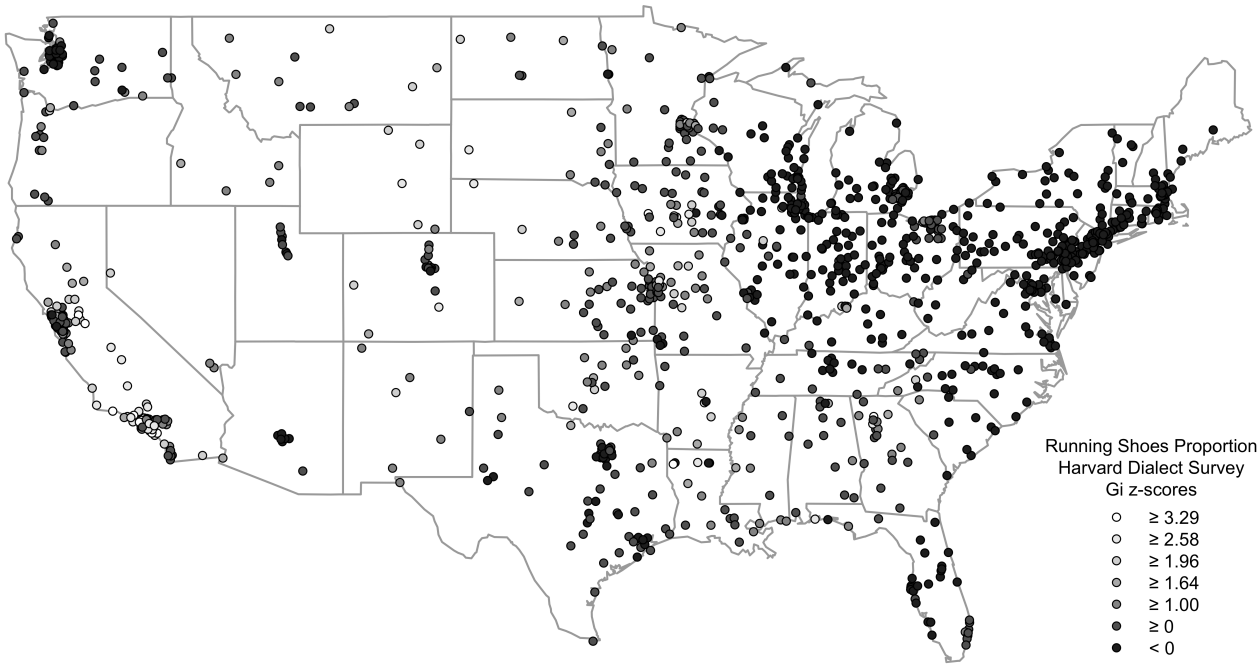
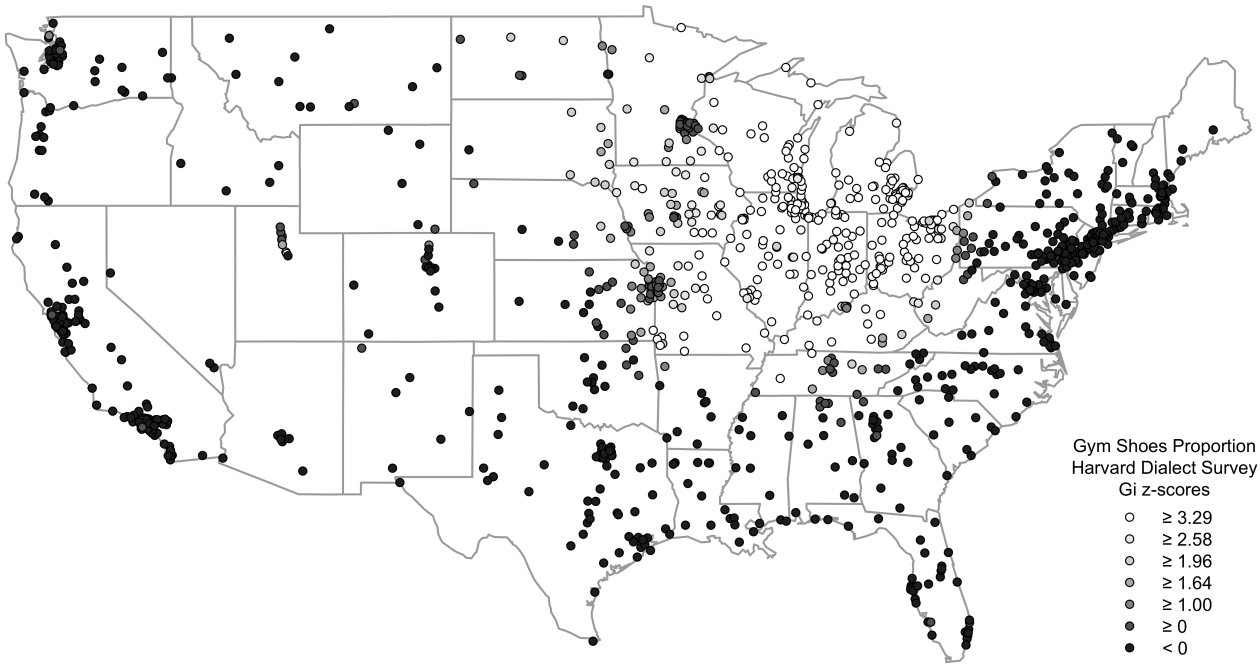


Figure 12 Local Autocorrelation Map for *Gym Shoes* (Harvard Dialect Survey)



5.2.2 Frosting/Icing

The second variable analyzed here is the alternation between *frosting* and *icing*. *Icing* is the most common variant accounting for 80% of the total hits. The proportion of *frosting* to *icing* was calculated for 934 cities and subjected to a local spatial autocorrelation analysis. In this case, because there are only two variants, it was only necessary to calculate and map one proportion because the proportion of the second variant is the exact inverse of the first. The local spatial autocorrelation map is presented in Figure 13. Unlike the maps presented above, this map gives equal weight to both positive and negative Getis-Ord *Gi z*-scores, with positive values (i.e. lighter dots) identifying clusters associated with the first variant (in this case *frosting*) and with negative values (i.e. darker dots) identifying clusters associated with the second variant (in this case *icing*). Grey dots represent regions of variability. This map shows that *frosting* is most common in the North, especially in the North Central States, and to a lesser extent in southeastern New England, and that *icing* is most common in the Southeast. The West and the Northeast are identified as regions of variability.

This analysis is largely confirmed by the HDS data. The local spatial autocorrelation map based on the proportion of HDS informants who prefer *frosting* to *icing* in 711 cities is presented in Figure 14 (excluding informants who said that they use both forms or that the two forms do not mean the same thing). The HDS map aligns closely with the map based on the data gathered through site-restricted web searches, aside from the West Coast, which is identified as a *frosting* region in the HDS map.

Figure 13 Local Autocorrelation Map for *Frosting/Icing* (Web Searches)

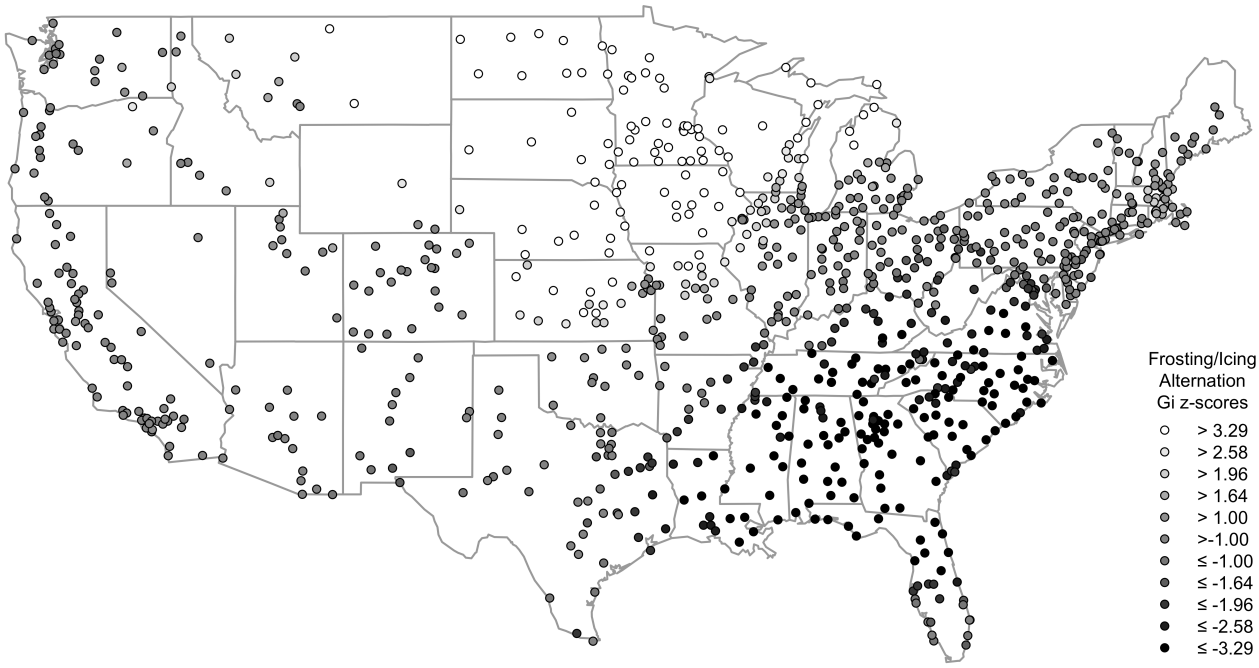
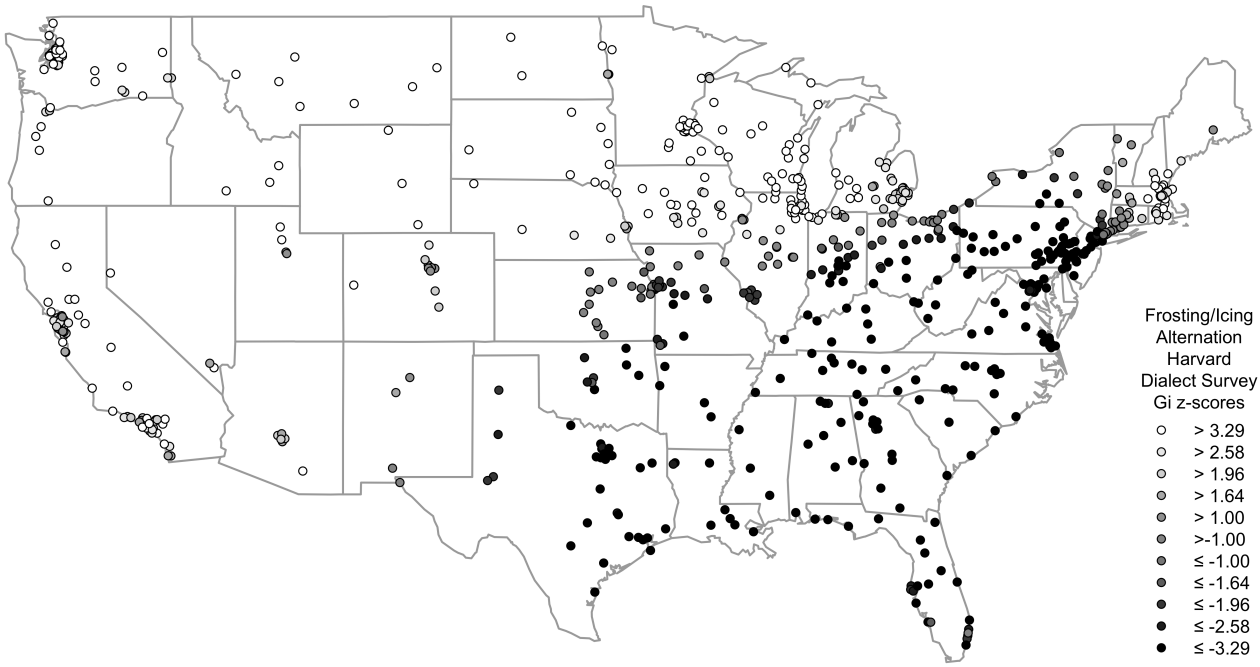


Figure 14 Local Autocorrelation Map for *Frosting/Icing* (HDS)



5.2.3 Trash Can/Garbage Can

The third variable analyzed here is the alternation between *trash can* and *garbage can*. The variants *waste basket* and *rubbish bin* were ignored because they are very infrequent. *Trash can* is the most common variant accounting for 59% of the total hits. The proportion of *trash can* to *garbage can* was calculated for 703 cities and subjected to a local spatial autocorrelation analysis. The local autocorrelation map is presented in Figure 15, showing that *garbage can* is most common in the Southeast and *trash can* is most common in the North. The Midland, the Lower Midwest and the Southwest are identified as regions of variability.

This analysis is confirmed by the HDS data. The local spatial autocorrelation map based on the proportion of HDS informants who prefer *trash can* to *garbage can* in 861 cities is presented in Figure 16. The HDS map aligns very closely with the map based on the data gathered through site-restricted web searches, including *garbage can* outliers in Texas, Florida, Georgia and Arizona, and the identification of New England as an area of transition. The main difference between the two maps is the lower peninsula of Michigan, which is identified as a strong *trash can* region in the HDS map but which is identified here as a region of transition.

Figure 15 Local Autocorrelation Map for *Trash Can/Garbage Can* (Web Searches)

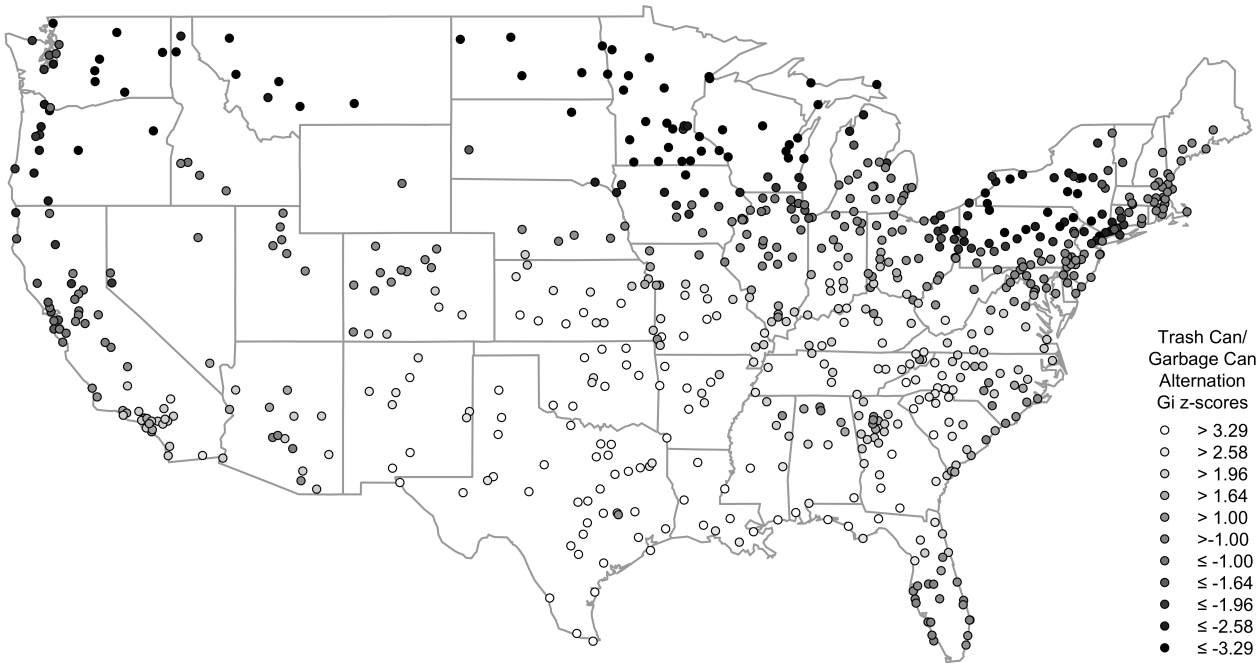
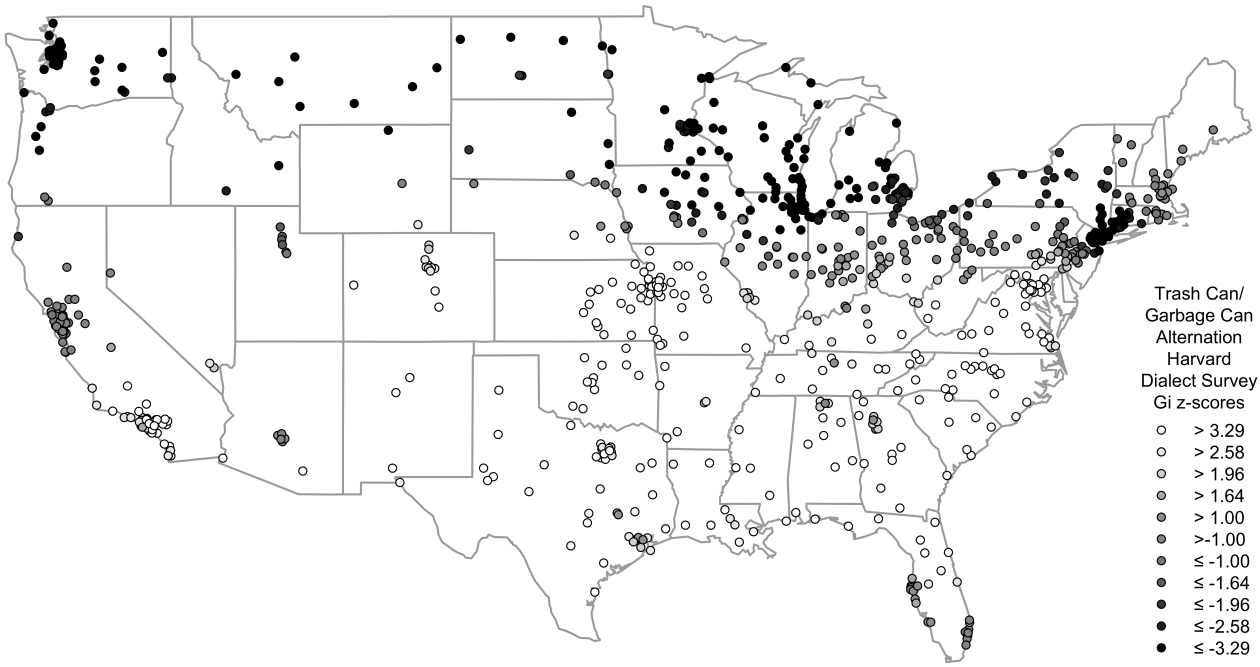


Figure 16 Local Autocorrelation Map for *Trash Can/Garbage Can* (HDS)



5.4.4 Drinking Fountain/Water Fountain

The fourth variable analyzed here is the alternation between *water fountain* and *drinking fountain.* The variant *bubbler* was ignored because it is both highly polysemous and very infrequent. *Water Fountain* is the most common variant accounting for 90% of the total hits. The proportion of *water fountain* to *drinking fountain* was calculated for 337 cities and subjected to a local spatial autocorrelation analysis. The local autocorrelation map is presented in Figure 17, showing that *water fountain* is most common in the eastern United States, especially in the Southeast, and *drinking fountain* is most common in the Midwest and the West.

This analysis is confirmed by the HDS data. The local autocorrelation map for the proportion of HDS informants who prefer *drinking fountain* to *water fountain* in 1,102 cities is presented in Figure 18. The HDS map aligns closely with the map based on the data gathered through site-restricted web searches, except that the pattern in the HDS map is stronger overall, especially the *drinking fountain* region, which stretches all the way to Texas.

Figure 17 Local Autocorrelation Map for *Water Fountain/Drinking Fountain* (Web Searches)

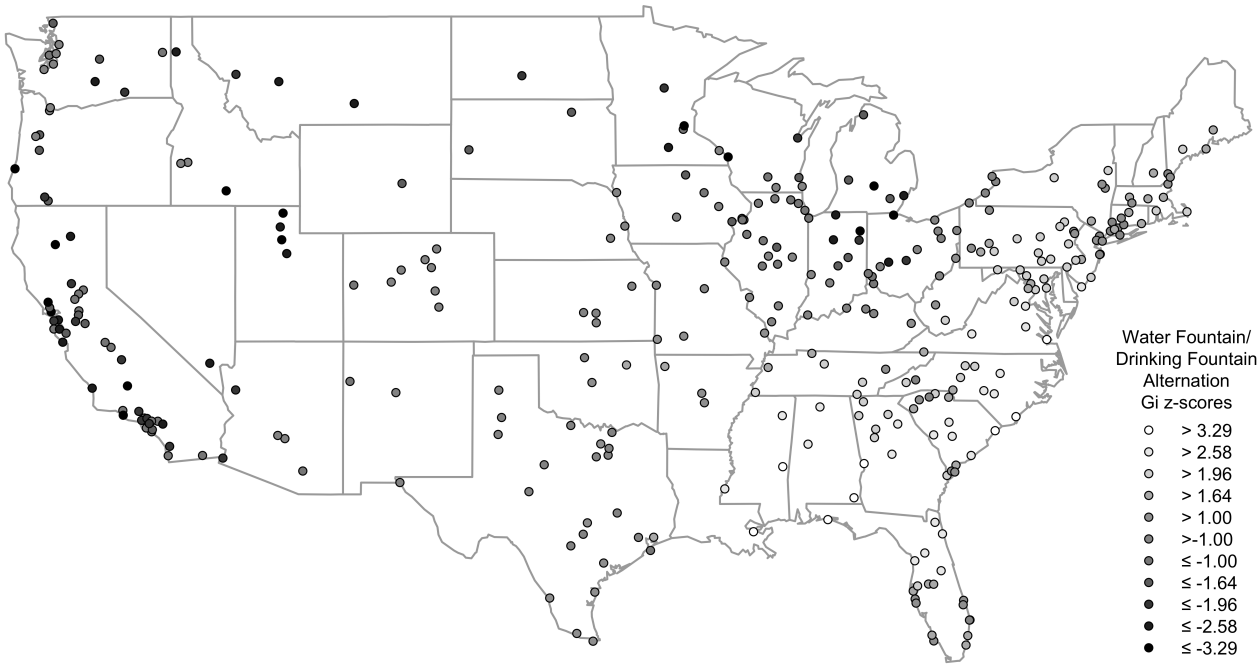
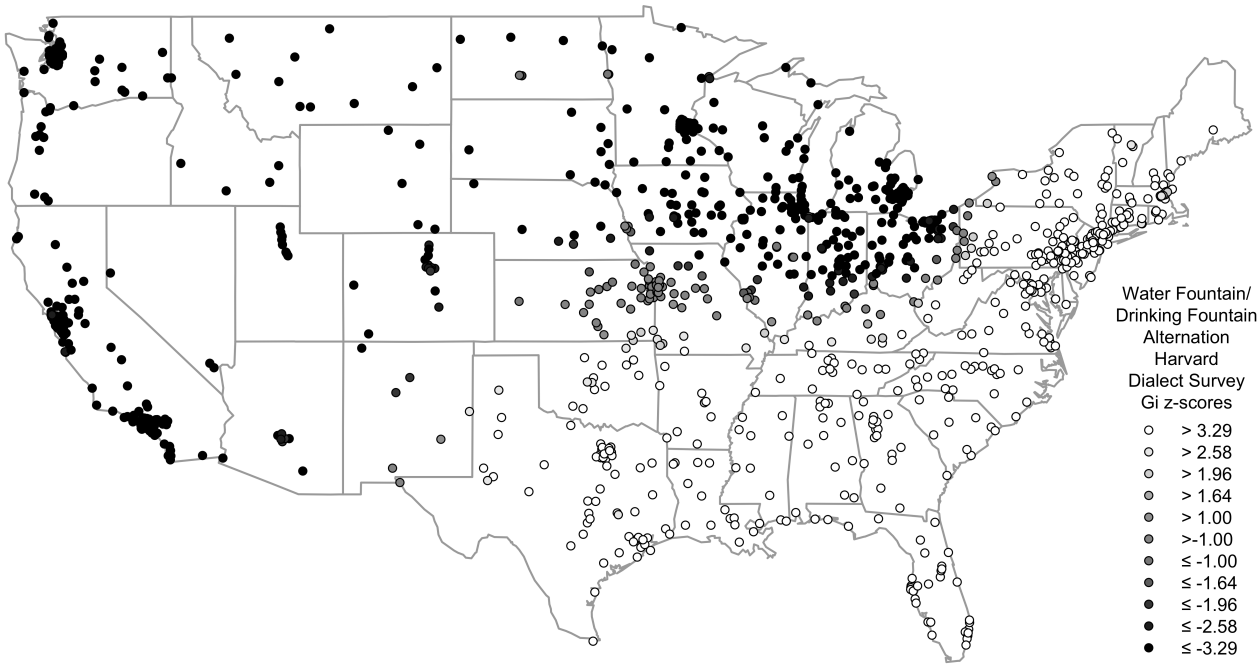


Figure 18 Local Autocorrelation Map for *Water Fountain/Drinking Fountain* (HDS)



5.4.5 Bag/Sack

The fifth variable analyzed here is the alternation between *bag* and *sack*. The variant *poke* was ignored because it is highly polysemous. *Bag* is the most common variant accounting for 82% of the total hits. The proportion of *bag* to *sack* was calculated for 1,217 cities and subjected to a local spatial autocorrelation analysis. The local autocorrelation map is presented in Figure 19, showing that *bag* is most common in the Northeast and the Mid Atlantic and *sack* is most common in the Central and South Central States. The West is identified as a region of variability.

This analysis is largely confirmed by the HDS data. The local spatial autocorrelation map for the proportion of HDS informants who prefer *bag* to *sack* in 1,146 cities is presented in Figure 20. The HDS map aligns with the map based on the data gathered through site-restricted web searches, except that the *bag* region stretches into the Midwest and California is also identified as a *bag* region. The *sack* region in the HDS map is also somewhat smaller, stronger and more homogeneous than the region identified here.

Figure 19 Local Autocorrelation Map for *Bag/Sack* (Web Searches)

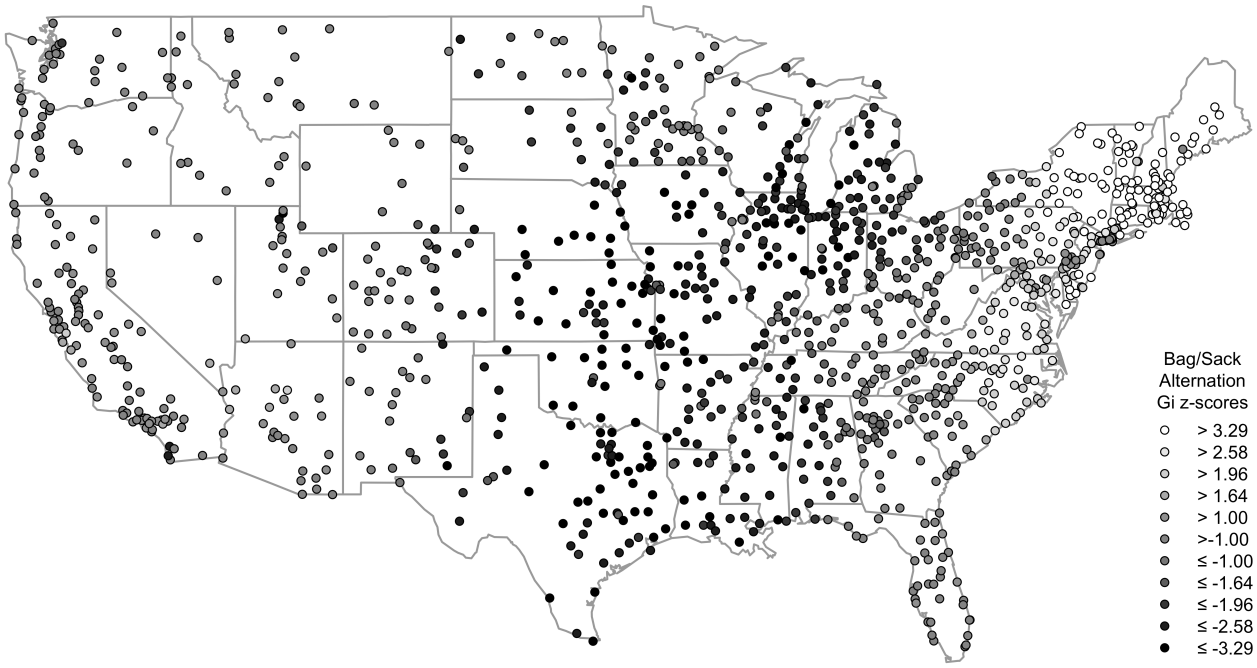
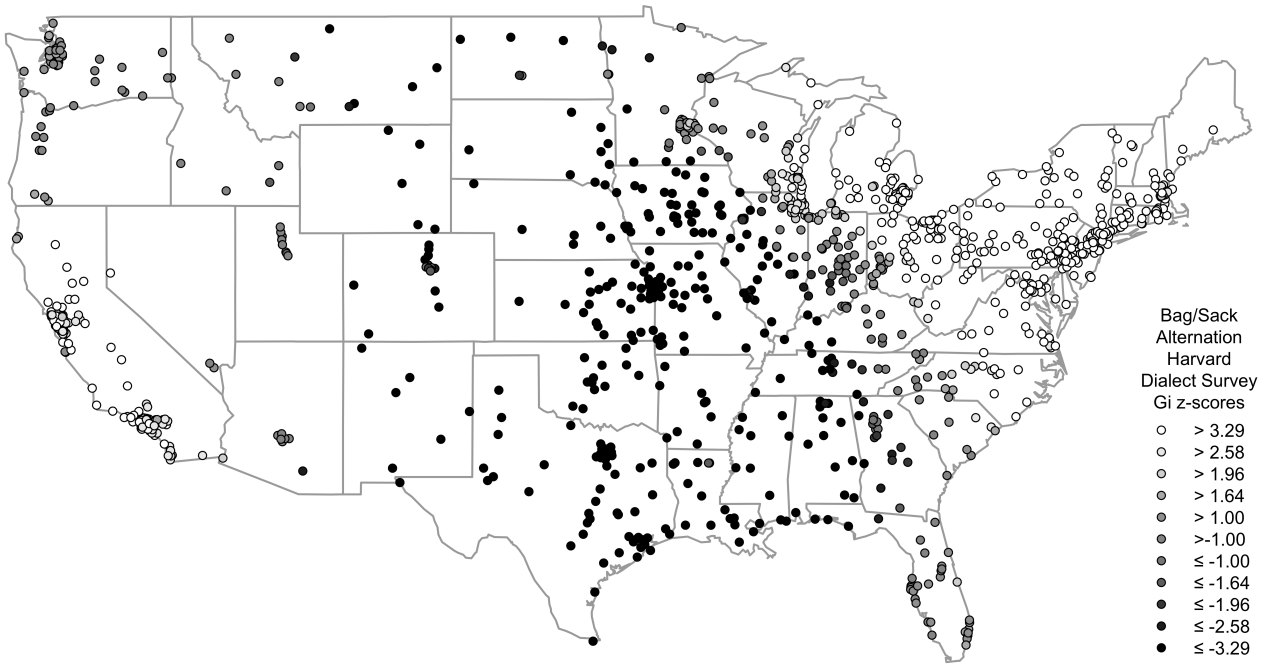


Figure 20 Local Autocorrelation Map for *Bag/Sack* (HDS)



5.4.6 Take-Out/Carry-Out

The sixth variable analyzed here is the alternation between *take-out* and *carry-out*. *Take-out* is the most common variant accounting for 52% of the total hits, although this term also appears to be considerably more polysemous than *carry-out*. The proportion of *take out* to *carry out* was calculated for 1,123 cities and subjected to a local spatial autocorrelation analysis. The local autocorrelation map is presented in Figure 21, showing that *take-out* is most common on the East and West Coast and *carry-out* is most common in the Central States and the Midwest. The South is identified as a region of variability.

This analysis is largely confirmed by the HDS data, despite the considerable amount of polysemy associated with *take-out*. The local spatial autocorrelation map for the proportion of HDS informants who prefer *take-out* to *carry-out* in 961 cities is presented in Figure 22. The HDS map aligns with the map based on the data gathered through site-restricted web searches, except that the *carry-out* region extends into the South Central States and the *take-out* region is larger and stronger in the West.

Figure 21 Local Autocorrelation Map for *Take Out/Carry Out* (Web Searches)

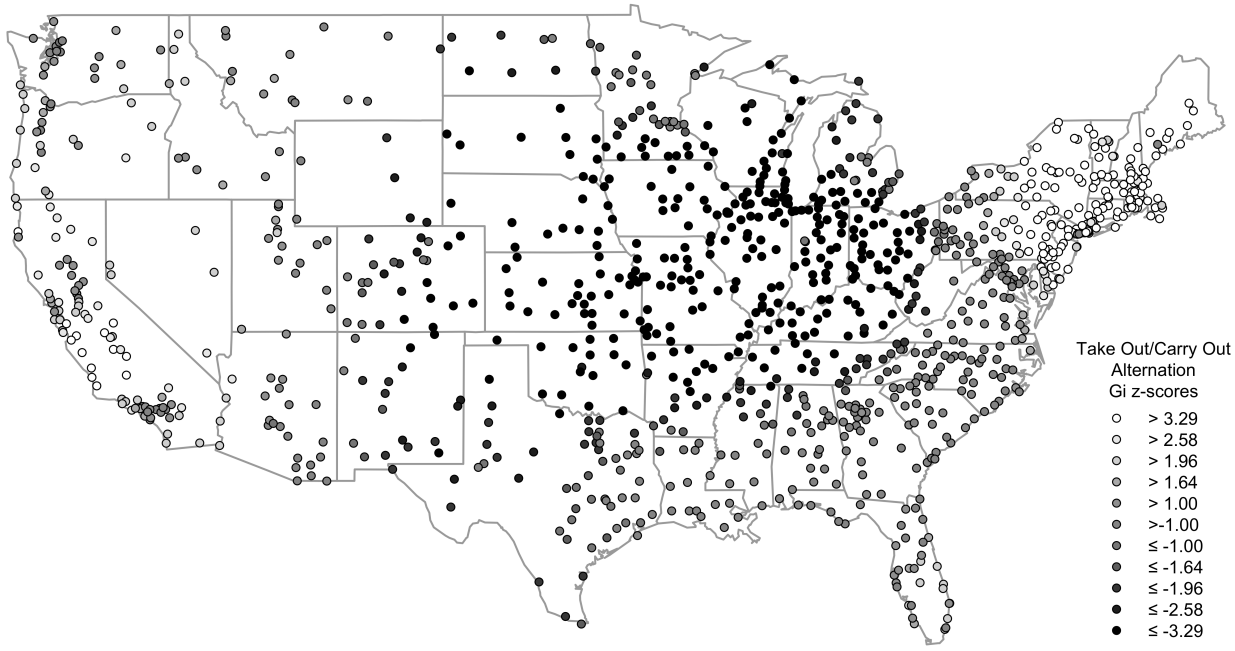
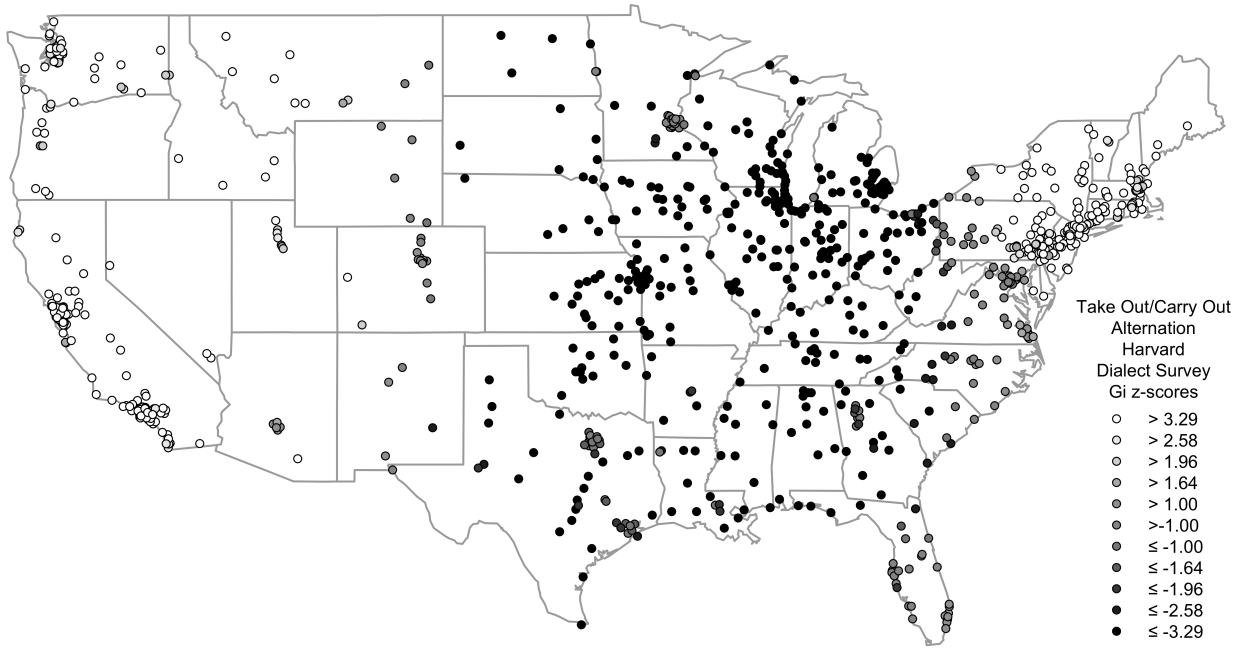


Figure 22 Local Autocorrelation Map for *Take Out/Carry Out* (HDS)



5.4.7 Cut the Grass/Mow the Lawn /Mow the Grass

The seventh variable analyzed here is the alternation between *mow the lawn, cut the grass* and *mow the grass*. The variant *cut the lawn* was ignored because it is very infrequent. *Cut the grass* is the most common variant accounting for 66% of the total hits, followed by *mow the lawn* accounting for 22% of the total hits and *mow the grass* accounting for 12% of the total hits. The proportions of the three variants were calculated for 273 cities and each set of proportions was subjected to a local autocorrelation analysis. The local autocorrelation maps for each of the three variants are presented in Figures 23-25, showing that *cut the grass* is most common in Southeast, excluding Florida, the Mid Atlantic States and upstate New York, *mow the lawn* is most common in the West, the Northern Great Plains and New England, and *mow the grass* is most common in the southern Midwest. Texas and Florida are identified as regions of variability.

This analysis is largely confirmed by the HDS data. The local autocorrelation maps for the proportions of HDS informants who prefer each of the three variants in 1,101 cities are presented in Figures 26-28. The HDS maps for the two most common variants align very closely with the maps based on the data gathered through site-restricted web searches. The HDS map for *mow the grass*, however, identifies a large region that encompasses the entire Southeast, whereas a smaller region that encompasses only the lower Midwest was identified here, which essentially constitutes an area of transition between the more common variants to the north and the south. In fact, the HDS map for *mow the grass* is very similar to the maps for *cut the grass* based on the data from both surveys, except that the *mow the grass* region does not extend as far into the Northeast and it extends father west into Texas and the Central States.

Figure 23 Local Autocorrelation Map for *Cut the Grass* (Web Searches)

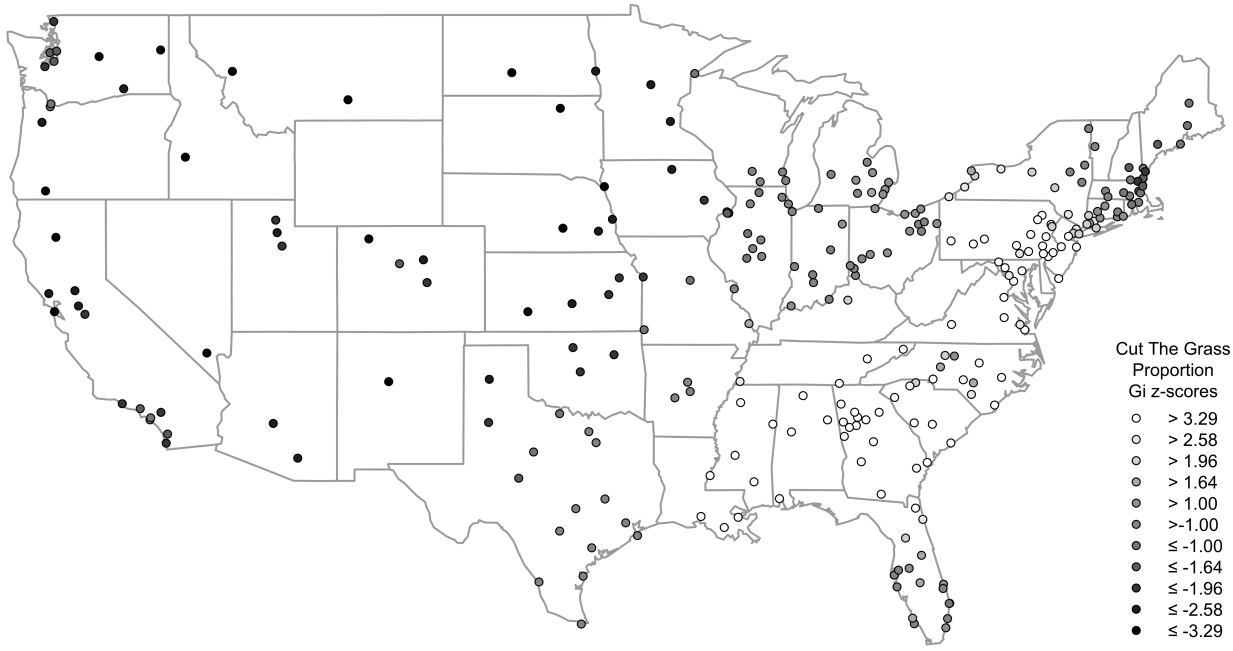


Figure 24 Local Autocorrelation Map for *Mow the Lawn* (Web Searches)

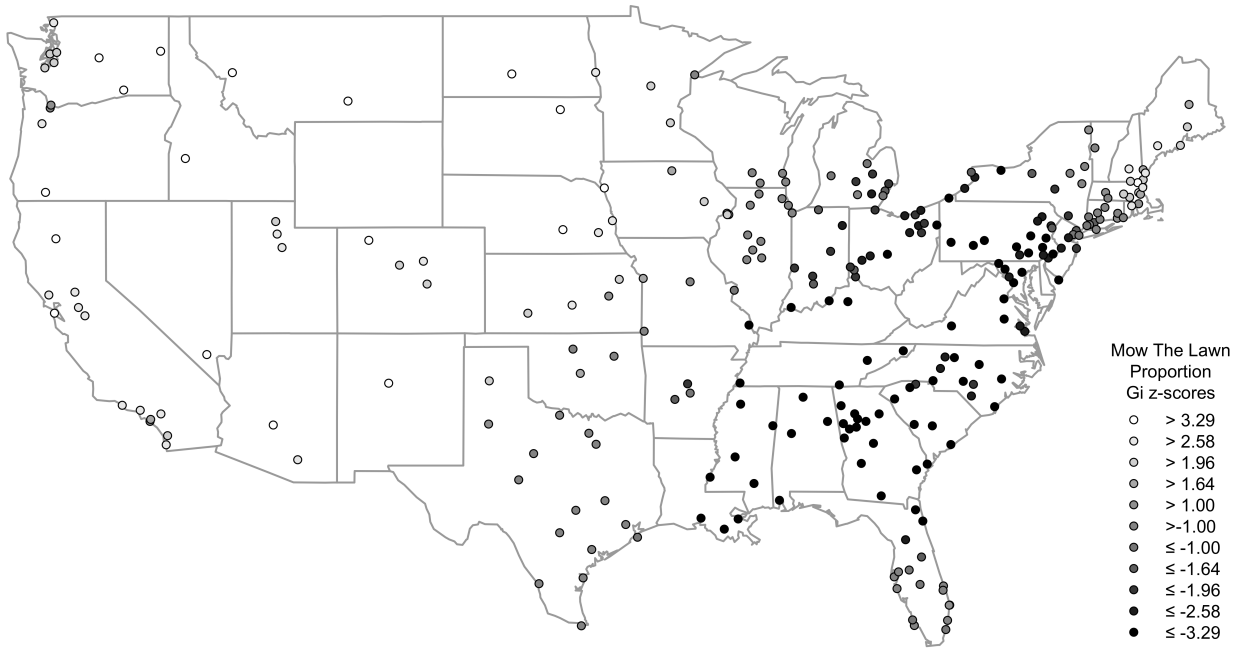


Figure 25 Local Autocorrelation Map for *Mow the Grass* (Web Searches)

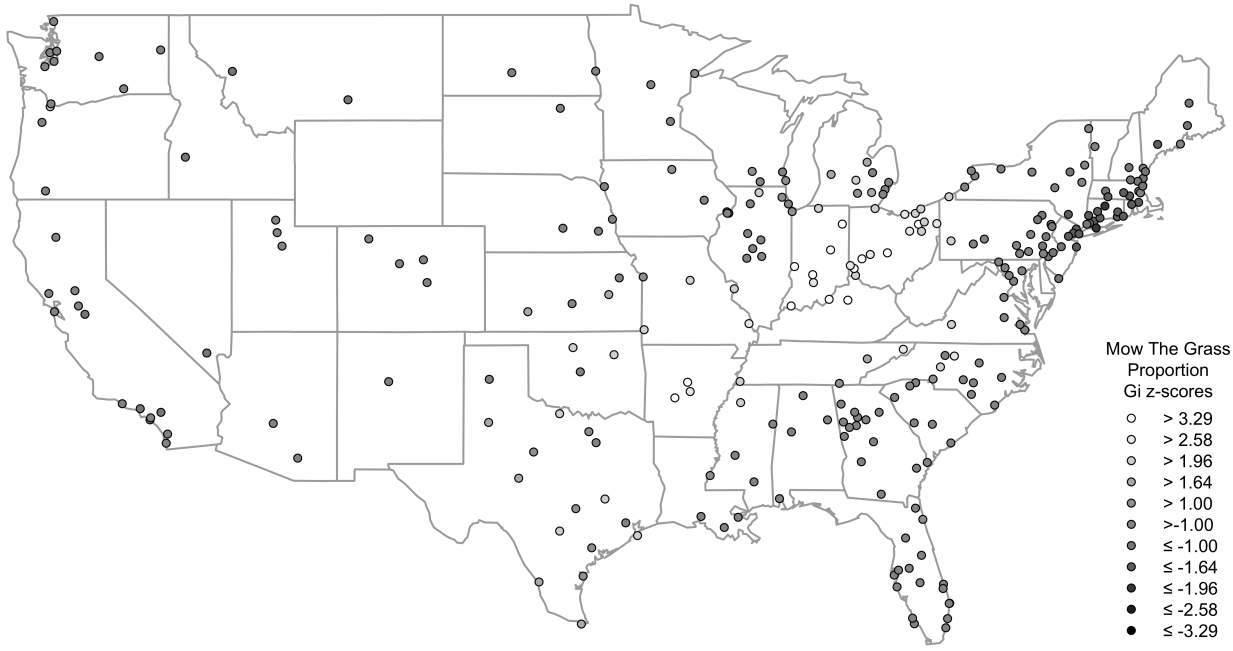


Figure 26 Local Autocorrelation for Map *Cut the Grass* (HDS)

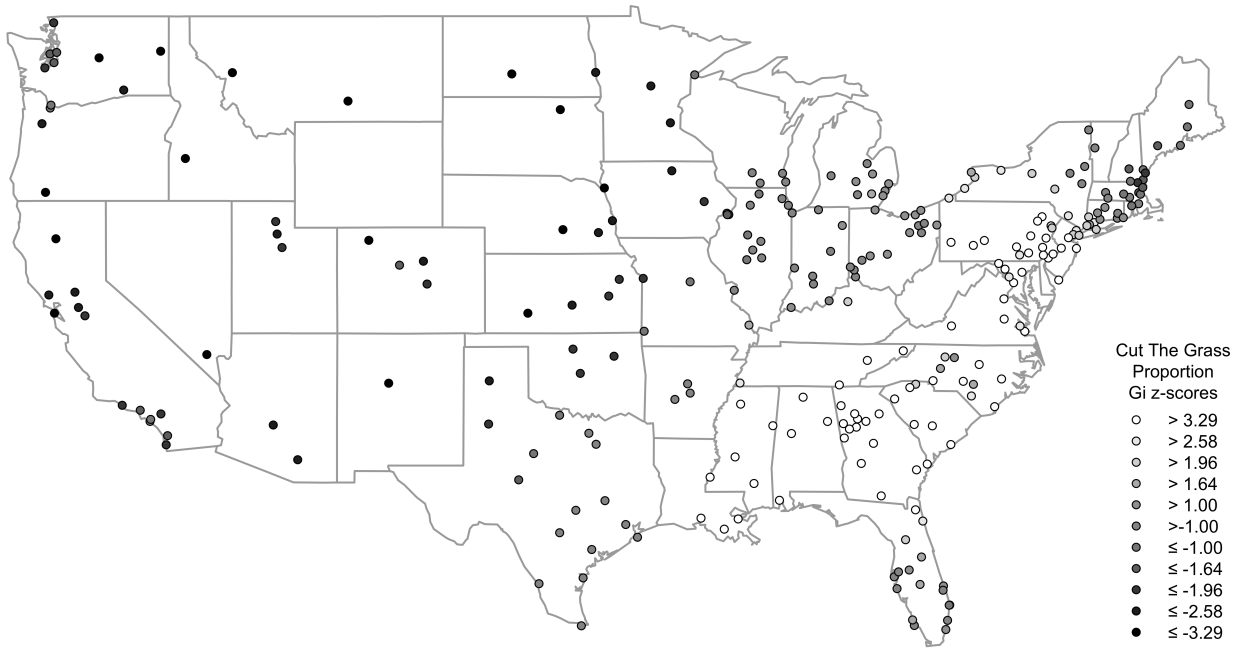


Figure 27 Local Autocorrelation Map for *Mow the Lawn* (HDS)

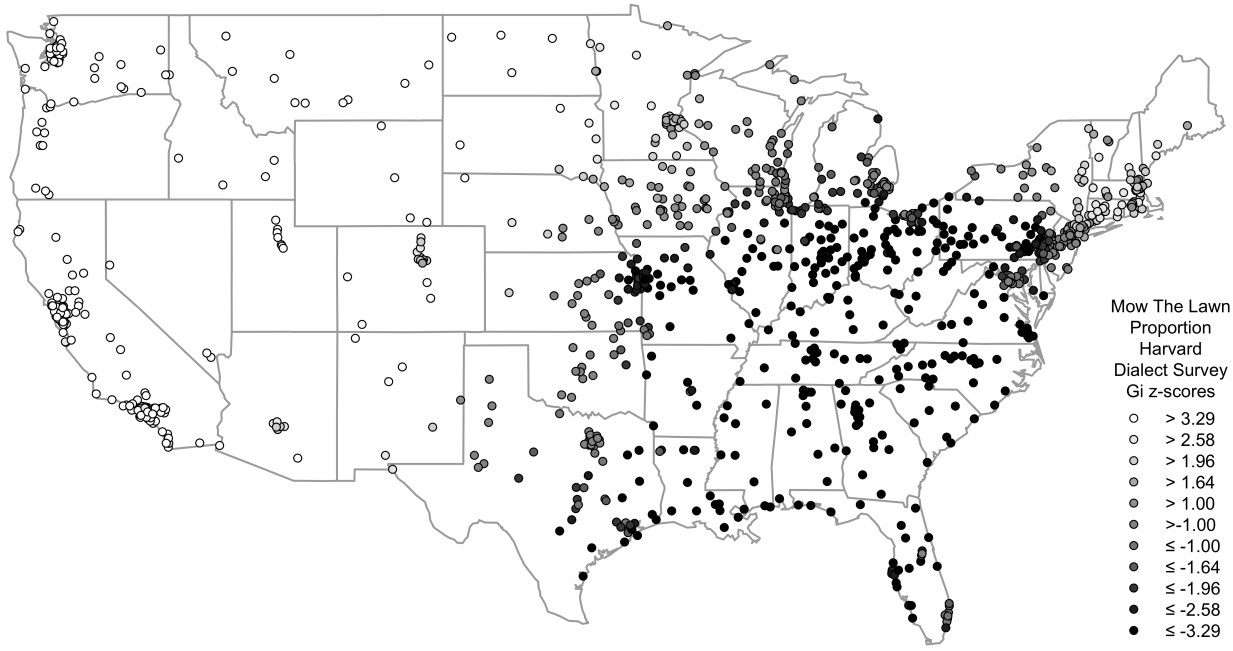
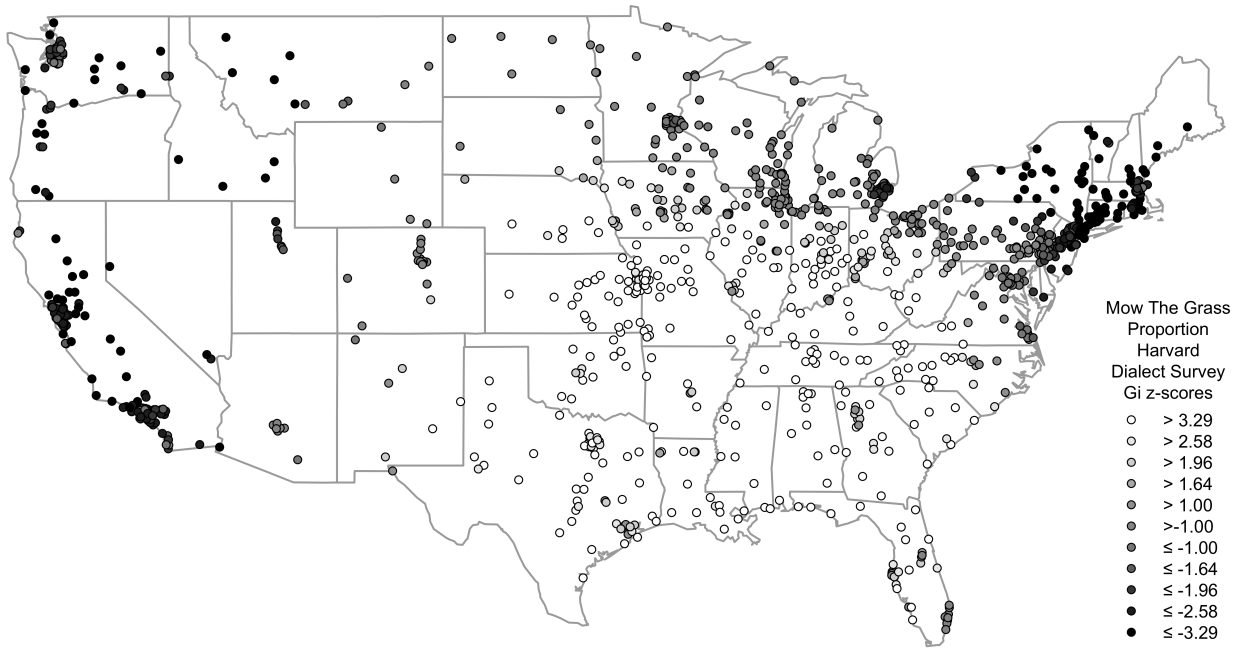


Figure 28 Local Autocorrelation Map for *Mow the Grass* (HDS)



5.4.8 Garage Sale/Yard Sale /Rummage Sale/Tag Sale

The eighth variable analyzed here is the alternation between *garage sale, yard sale, rummage sale,* and *tag sale.* Other variants including *stoop sale* and *thrift sale* were ignored because they are very infrequent. *Garage sale* is the most common variant accounting for 57% of the total hits, followed by *yard sale* accounting for 22% of the total hits, *rummage sale* accounting for 6% of the total hits, and *tag sale* accounting for 3% of the total hits. The proportions of the four variants were calculated for 1,122 cities and each set of proportions was subjected to a local autocorrelation analysis. The local autocorrelation maps for each of the four variants are presented in Figures 29-32, showing that *garage sale* is most common in the Midwest, the Central States and the Pacific Northwest, *yard sale* is most common in the Eastern United States, *rummage sale* is most common in the North, and *tag sale* is most common in Western New England. The Southwest was identified as a region of variability.

This analysis is confirmed by the HDS data. The local autocorrelation maps for the proportion of HDS informants who prefer each of the four variants in 1,130 cities are presented in Figures 33-36. These HDS maps align closely with the maps based on the data gathered through site-restricted web searches, except that California is identified as a *garage sale region* in the HDS map but as *rummage sale* region here.

Figure 29 Local Autocorrelation Map for *Garage Sale* (Web Searches)

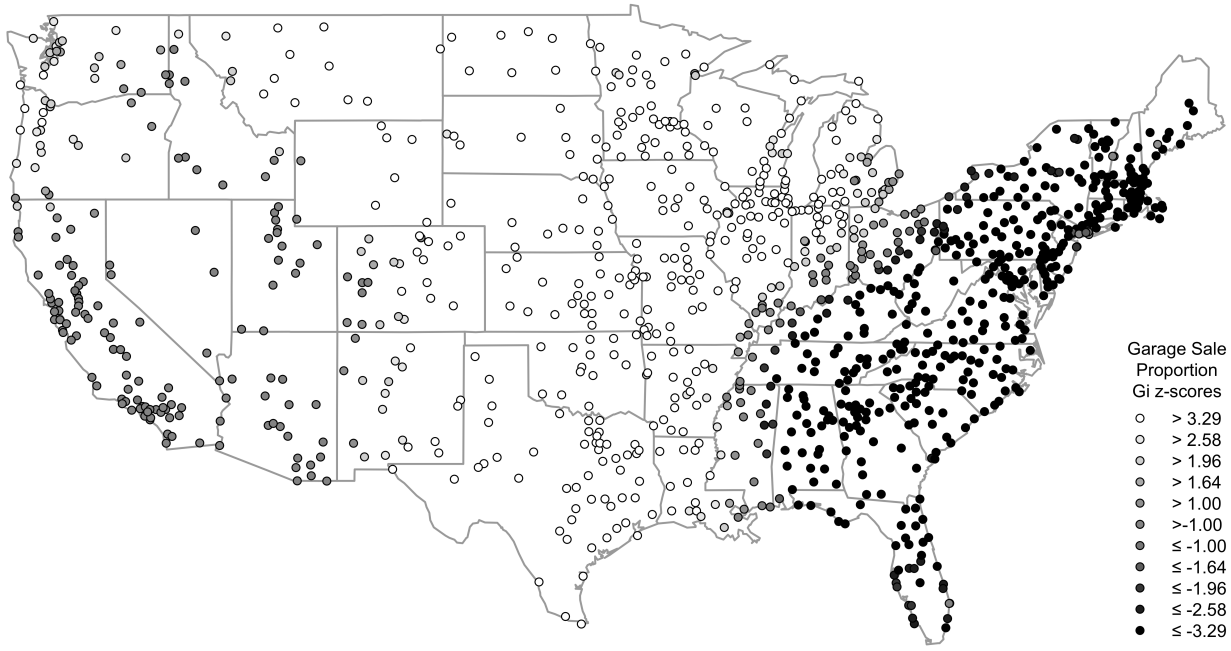


Figure 30 Local Autocorrelation Map for *Yard Sale* (Web Searches)

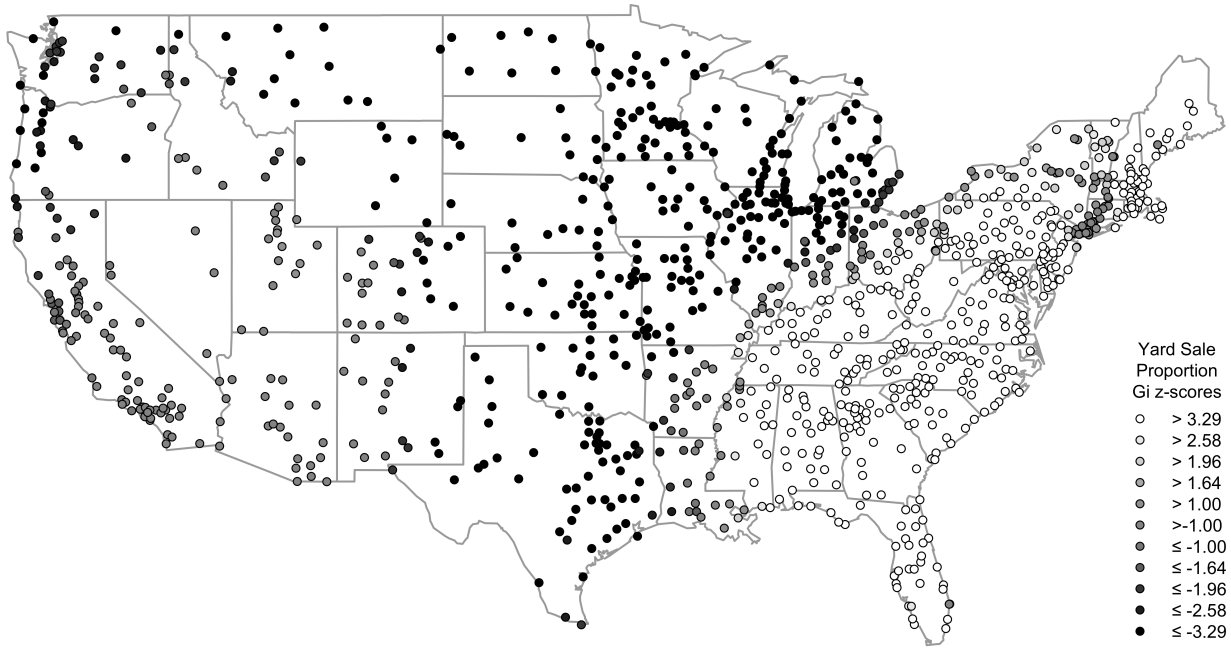


Figure 31 Local Autocorrelation Map for *Rummage Sale* (Web Searches)

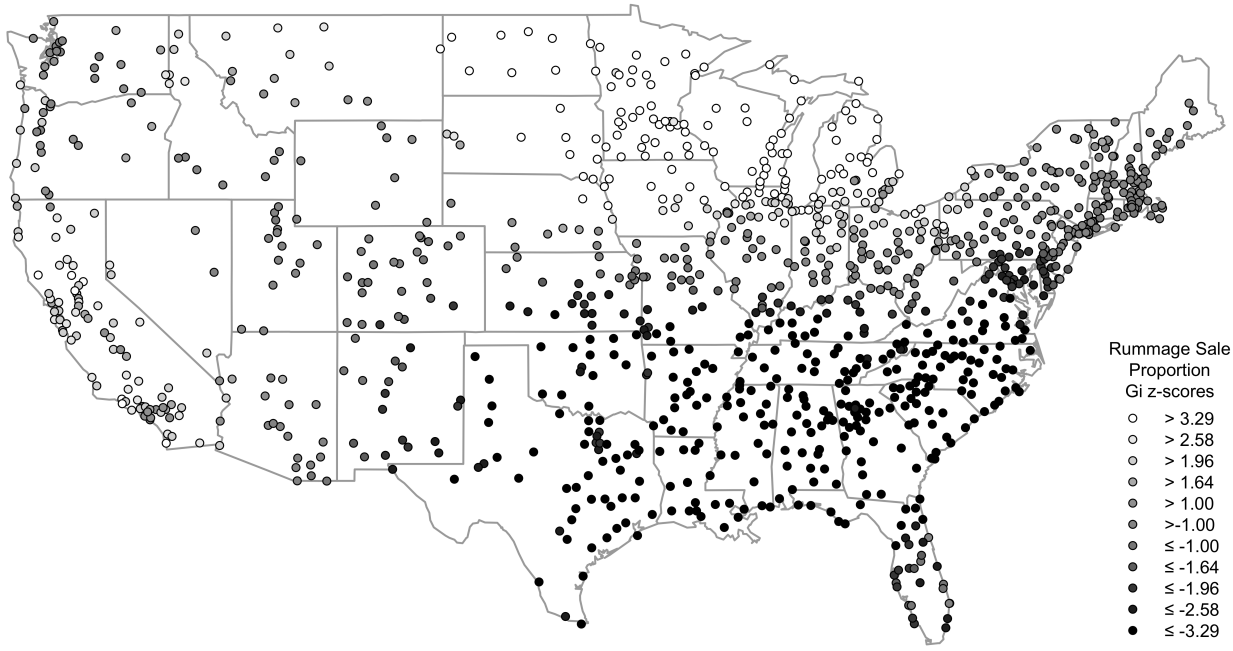


Figure 32 Local Autocorrelation Map for *Tag Sale* (Web Searches)

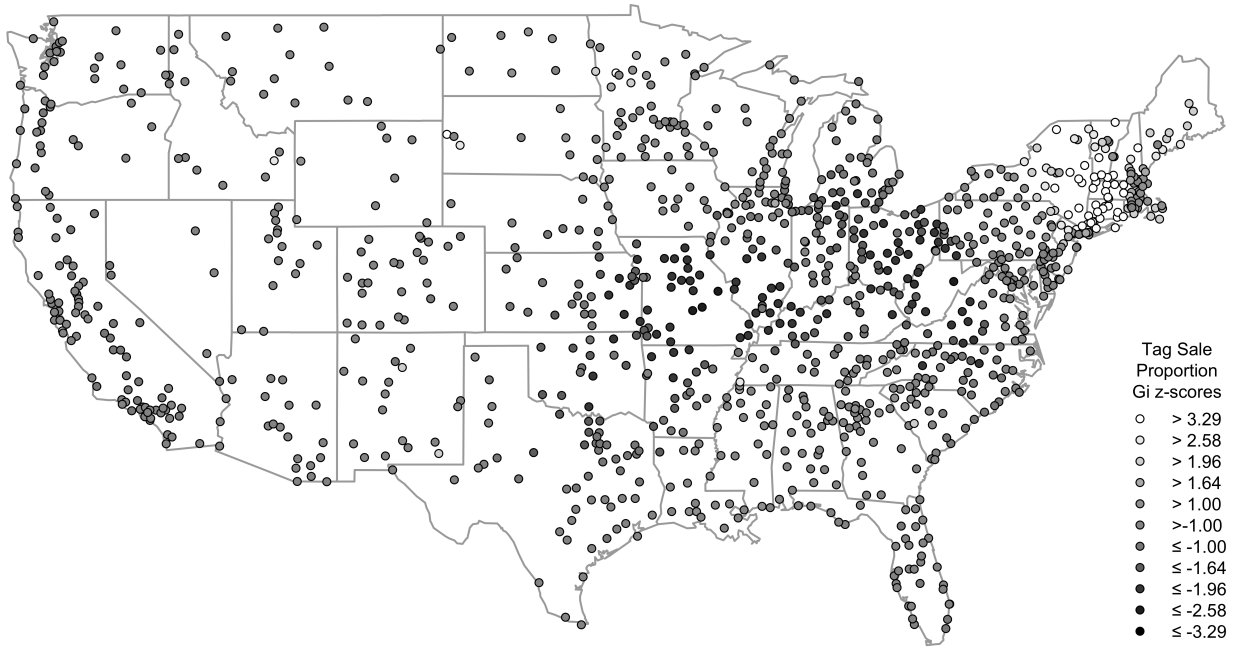


Figure 33 Local Autocorrelation Map for *Garage Sale* (HDS)

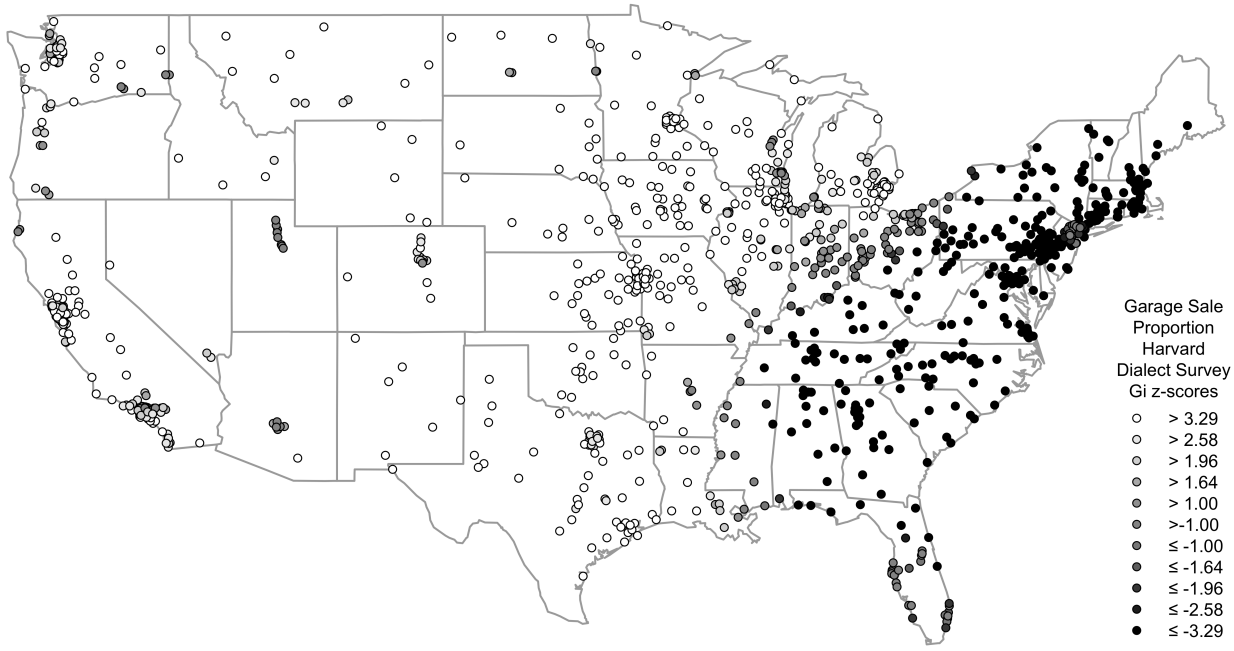


Figure 34 Local Autocorrelation Map for *Yard Sale* (HDS)

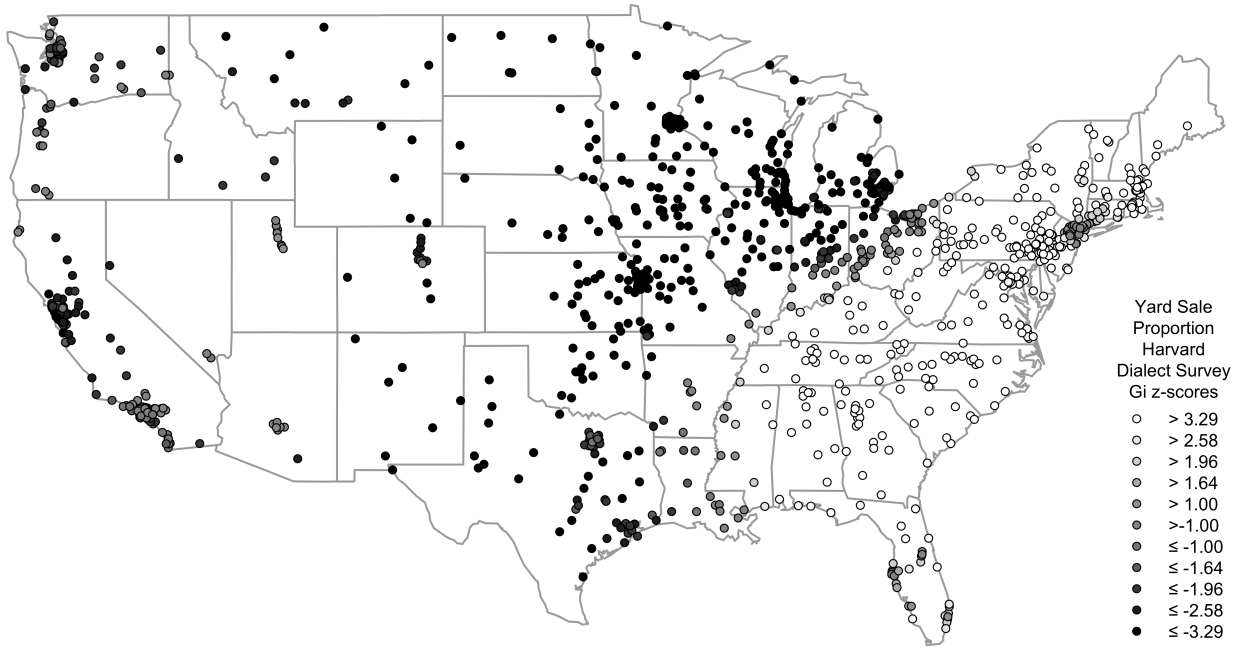


Figure 35 Local Autocorrelation Map for *Rummage Sale* (HDS)

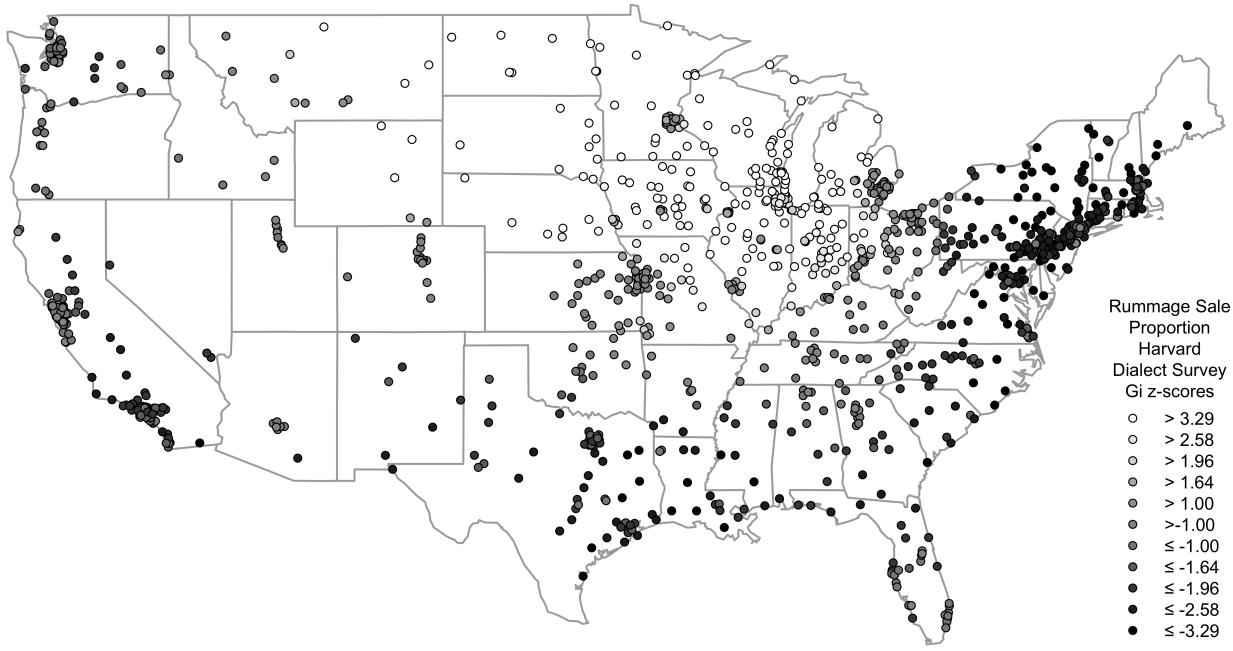
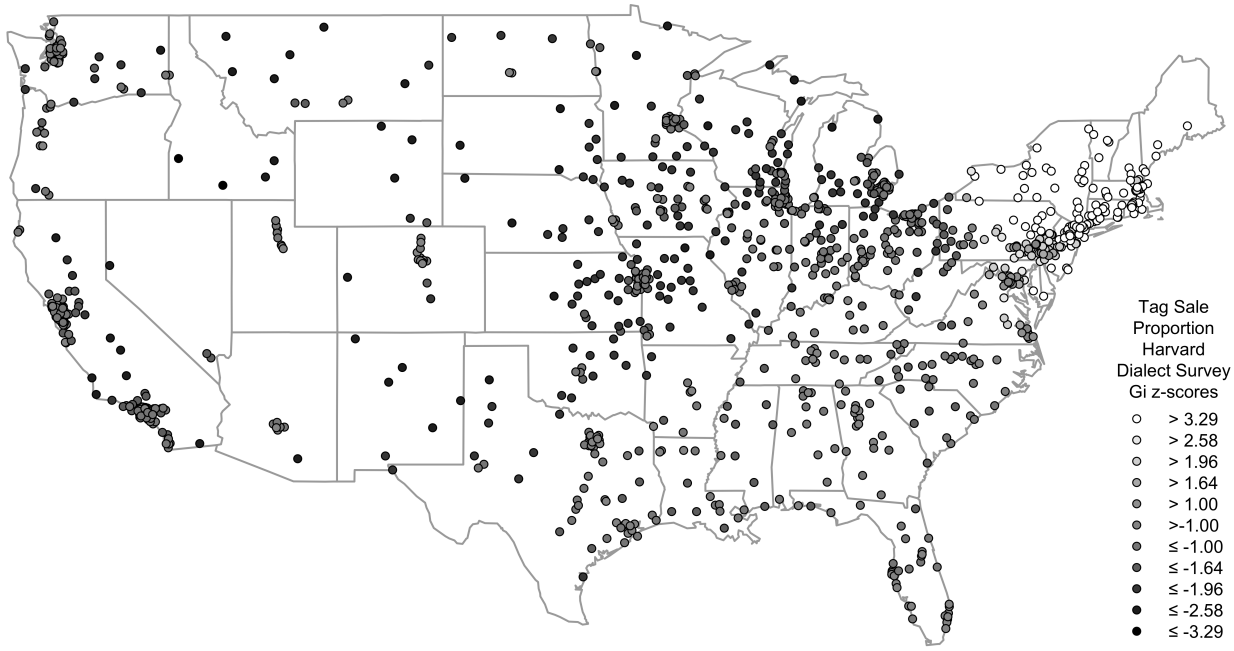


Figure 36 Local Autocorrelation Map for *Tag Sale* (HDS)



5.4.9 Grandmother/Granny/Grandma/Nana

The ninth and final variable analyzed here is the alternation between *grandmother, granny, grandma,* and *nana*. The variant *grammy* was ignored because it is highly polysemy (i.e. the music award). *Grandmother* is the most common variant accounting for 74% of the total hits, followed by *granny* accounting for 14% of the total hits, *grandma* accounting for 8% of the total hits, and *nana* accounting for 4% of the total hits. The proportions of the four variants were calculated for 1,205 cities and each set of proportions was subjected to a local autocorrelation analysis. The local autocorrelation maps for each of the four variants are presented in Figures 37-40, showing that *grandmother* is most common on the East Coast and in the Southeast, *granny* is most common the Deep South and the southern Midwest, *grandma* is most common in the Midwest and in the northern half of the West, and *nana* is most common in the Southwest.

This analysis is largely confirmed by the HDS data. The local autocorrelation maps for the proportions of HDS informants who prefer each of the four variants in 816 cities are presented in Figures 41-44, based on the combined counts for both maternal and paternal grandmothers, which were two separate items in the HDS. These HDS maps do not align perfectly with the maps based on the data gathered through site-restricted web searches. In particular, the *grandmother* region in the HDS map does not include the Northeast and the *nana* region in the HDS map is in the Northeast as opposed to the Southwest, which presumably reflects the large Hispanic population in this region. These differences are likely due to differences in data collection. The HDS asked informants which variant they used as a "nickname" for their grandparents, whereas it was only possible to search the web for the variants that are most common. This is why the standard form *grandmother* is by far the most frequent variant here, while only accounting for about 5% of the responses in the HDS.

Figure 37 Local Autocorrelation Map for *Grandmother* (Web Searches)

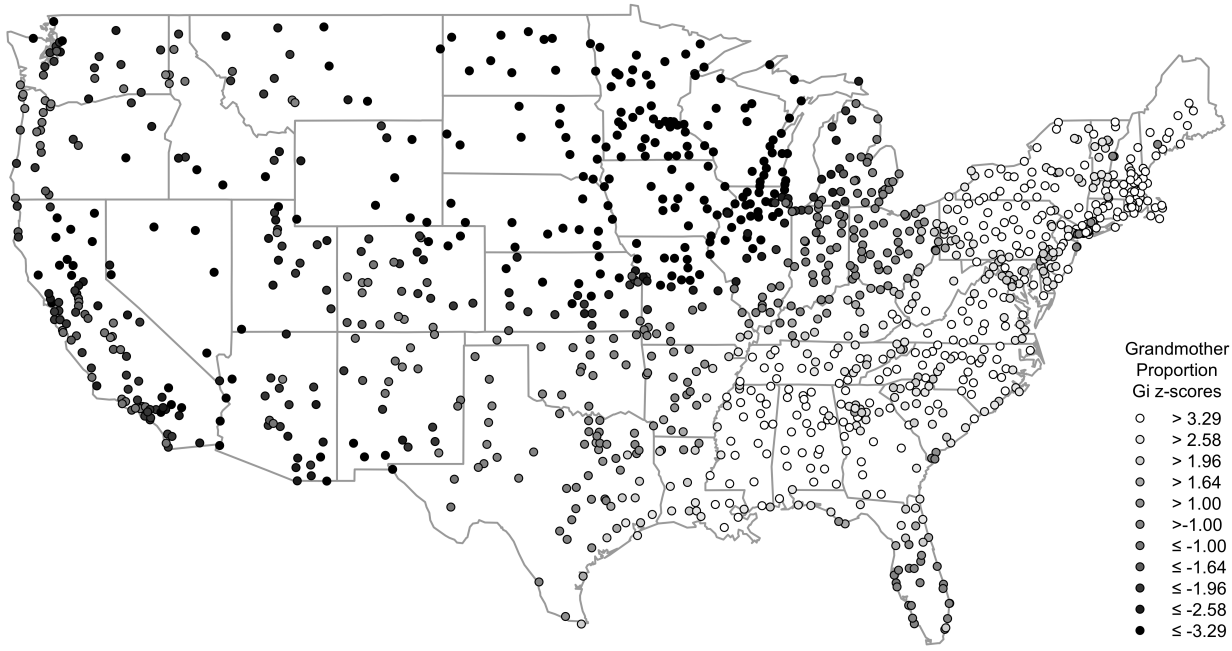


Figure 38 Local Autocorrelation Map for *Granny* (Web Searches)

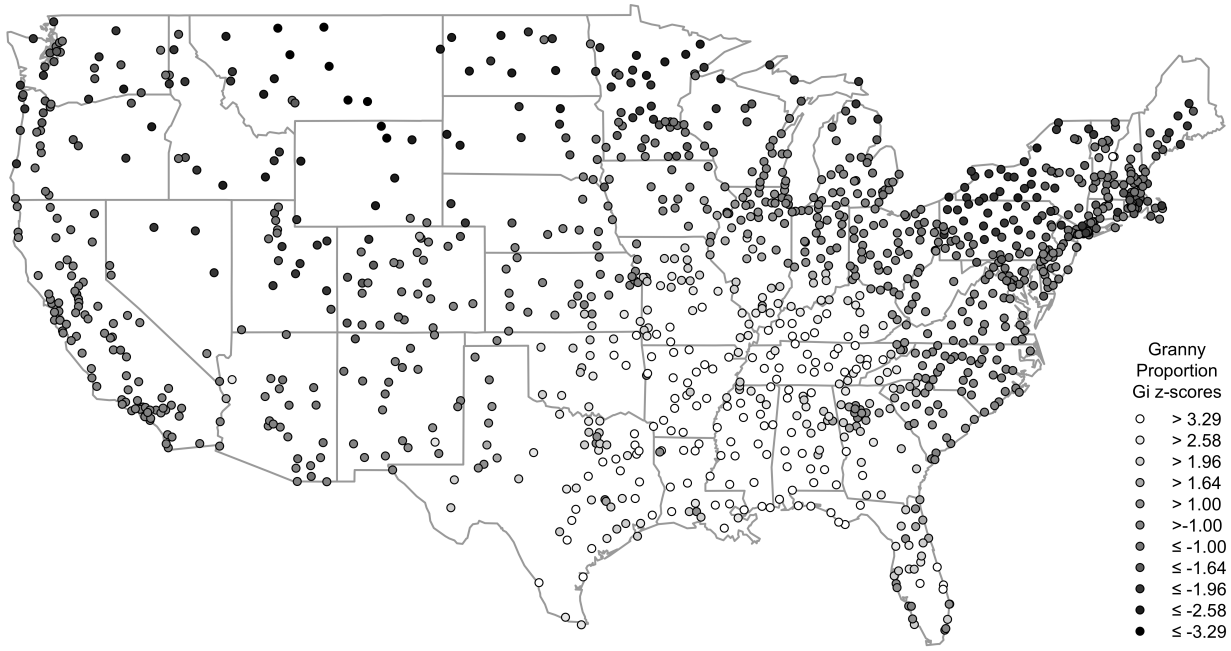


Figure 39 Local Autocorrelation Map for *Grandma* (Web Searches)

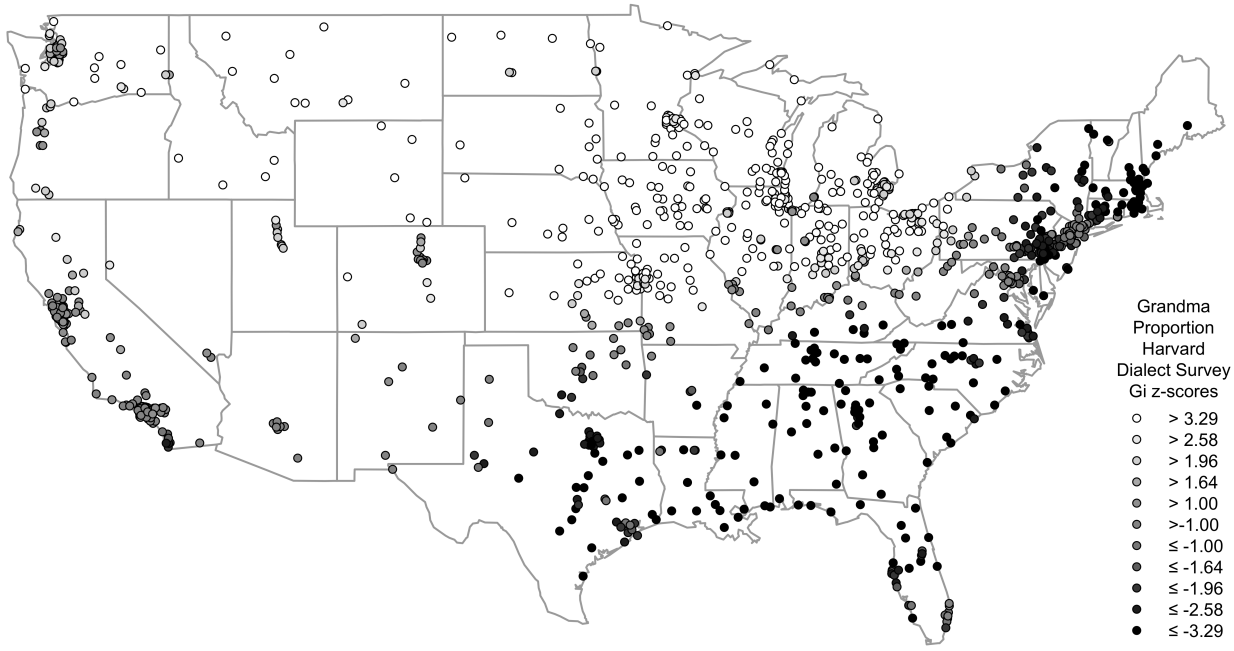


Figure 40 Local Autocorrelation Map for *Nana* (Web Searches)

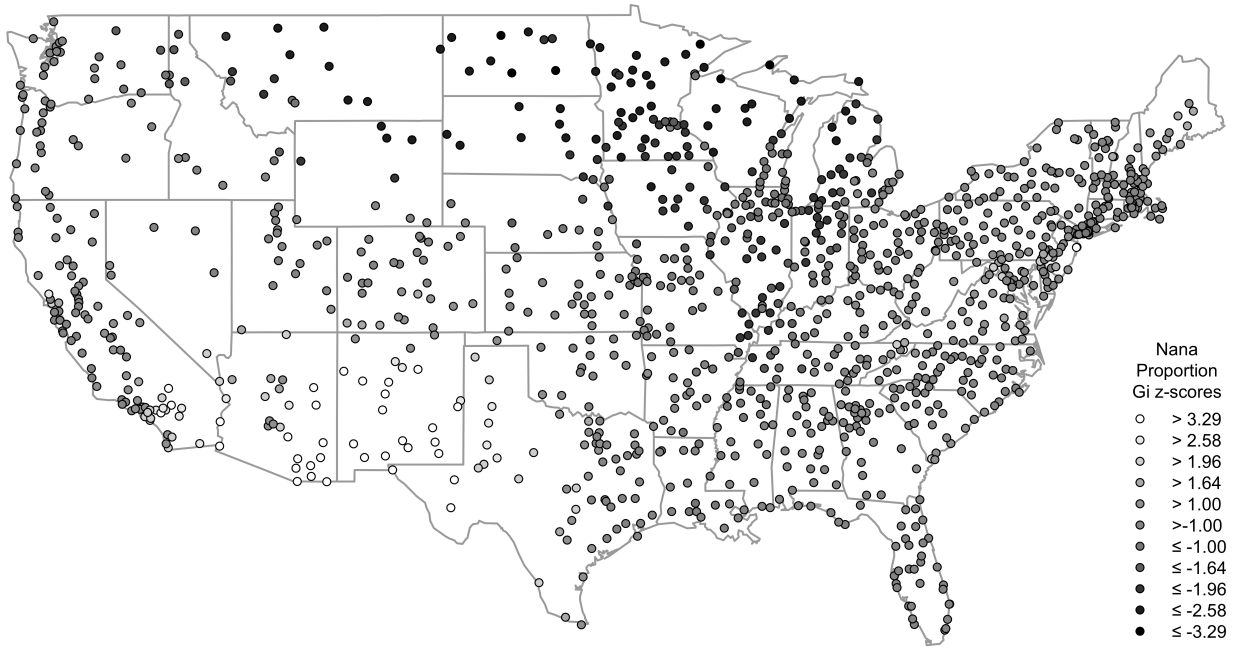


Figure 41 Local Autocorrelation Map for *Grandmother* (HDS)

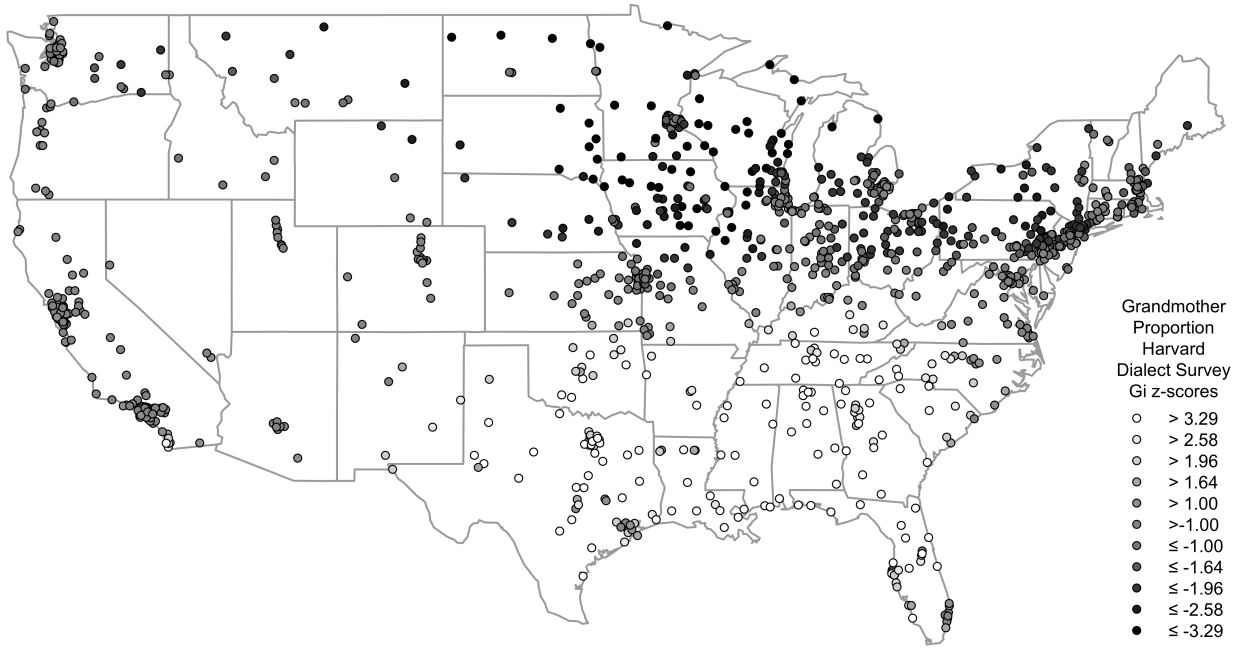


Figure 42 Local Autocorrelation Map for *Granny* (HDS)

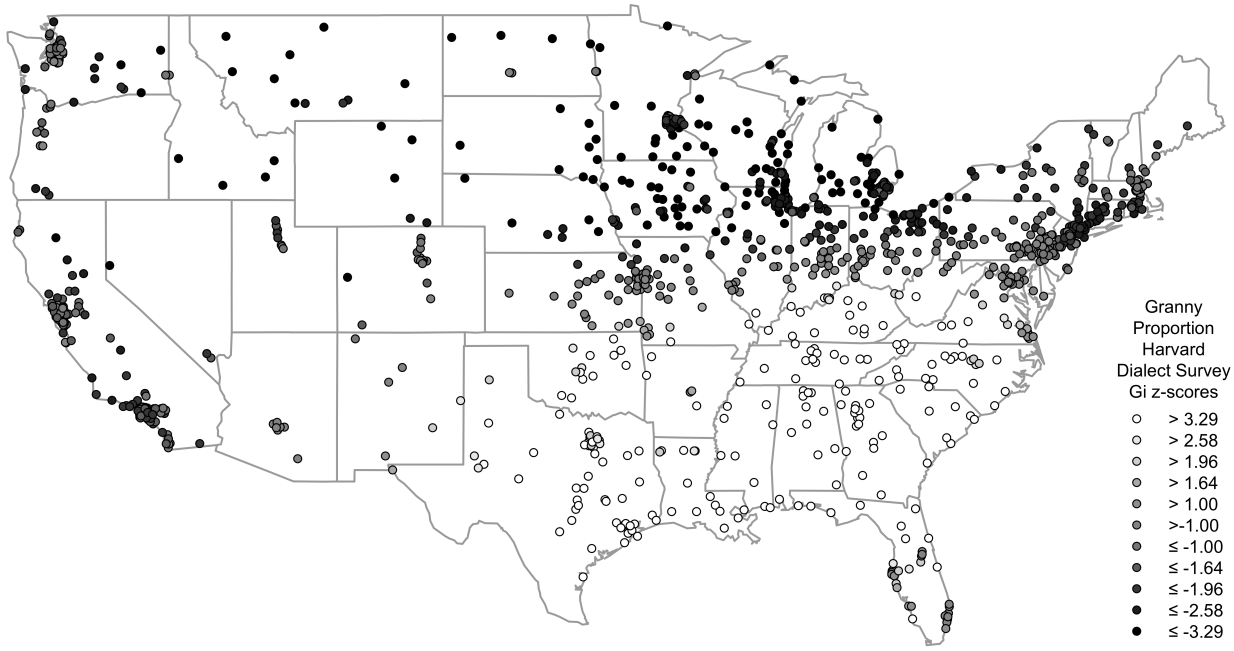


Figure 43 Local Autocorrelation Map for *Grandma* (HDS)

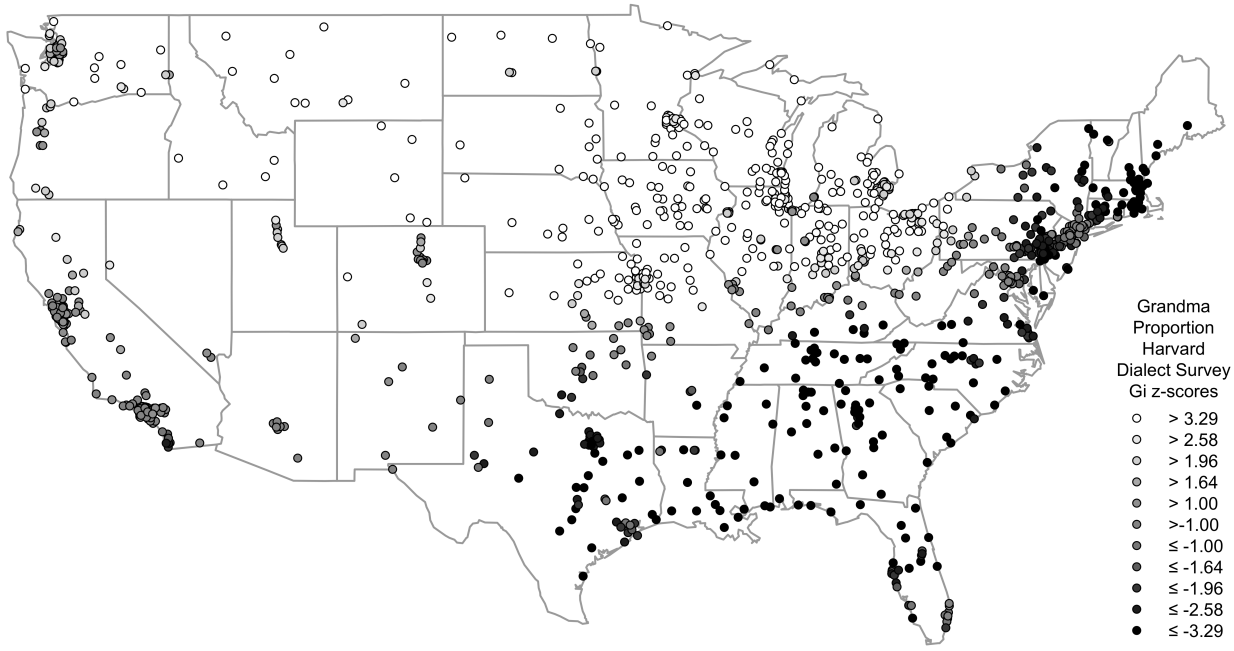
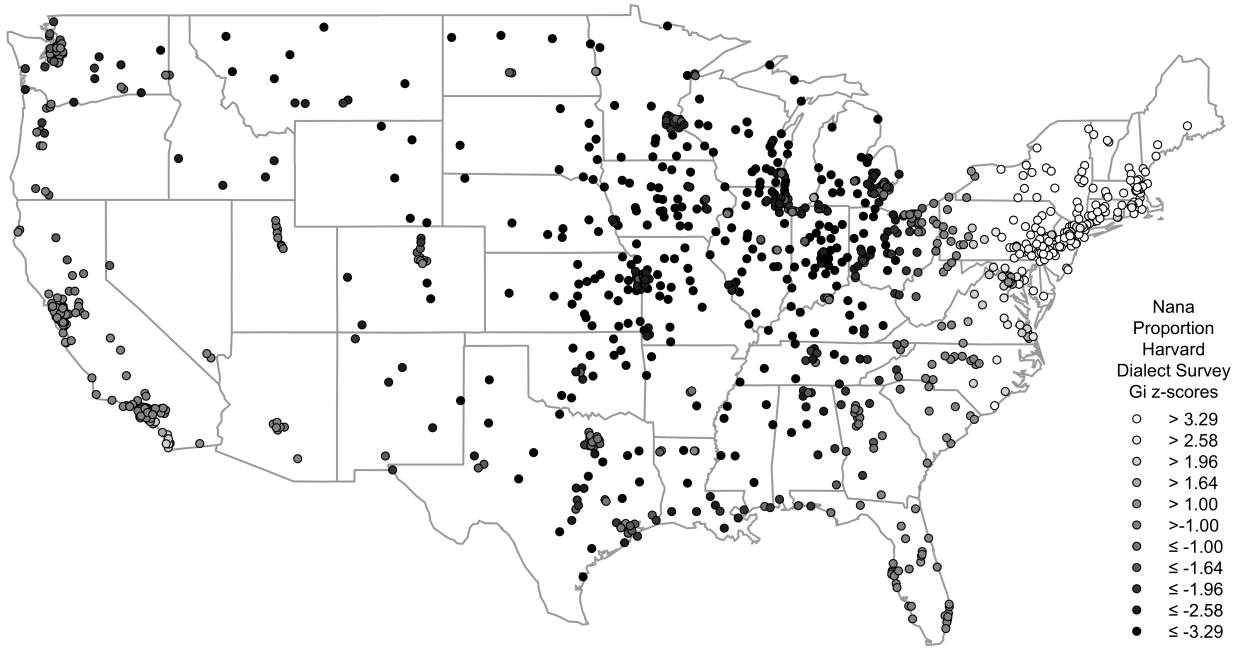


Figure 44 Local Autocorrelation Map for *Nana* (HDS)



6. Discussion

Overall, the maps generated through site-restricted web searches match the maps based on the data from the HDS remarkably well. There are certainly some differences between these maps, but in many of the cases the matches are almost perfect and in almost all of the cases the same basic patterns are present in both datasets. The only two maps that are very different, for the variants *mow the grass* and *nana*, involve the least frequent variants for those variables, and in fact the regions identified here seem plausible, as discussed above. Otherwise the differences between the two sets of maps are of degrees, involving slight shifts in the locations of the clusters—perhaps due only to differences in the locations, registers or eras that were sampled for the two surveys. The method has therefore been shown to be capable of accurately mapping lexical alternation variables in American English.

It would have been preferable to validate this method based on a larger set of content word alternations, but there are very few maps available for individual content word alternations in American English. In fact, there are almost no other content word alternations with known distributions in American English that can be used to evaluate the accuracy of the method for data collection being introduced here. The few remaining lexical variables from the HDS either involve variants that are too polysemous or too infrequent to be analyzed using this method, and as noted above, the *Dictionary of American Regional English*, which is the only other lexical survey to map content word alternations across the United States, does not provide proper maps or make its data available for individual variables, making it impossible to use this data to evaluate this method. This lack of data on content word alternations is precisely why a new method for collecting regional lexical data is required. It is very difficult to gather quantitative data on content word alternations using any other method: linguistic interviews are slow and expensive and it is difficult to observe a sufficient amount of language to get reliable frequency data for low frequency content word alternations using either linguistic interviews or a corpus-based approach to data collection.

Despite the promise of this method, it must be applied with care. Variables with highly polysemous variants are problematic, as are variables with variants that occur frequently as part of proper nouns or idiomatic expressions. When using site-restricted web searches to count the variants of

alternation variables, counting some non-interchangeable uses is unavoidable, but to some extent it is only another source of noise that can be overcome by large amounts of data and the application of appropriate statistics. However, when the non-interchangeable uses of variants are more common than the interchangeable uses, an alternation variable probably cannot be observed using the basic method being introduced here. When testing the method numerous variables with highly polysemous variants (e.g. *soda/pop* alternation) were analyzed unsuccessfully. It may be possible to measure variables like these by counting variants in specific contexts where they are generally interchangeable (e.g. *drink a soda/pop*) or by counting the variants in specific contexts where they are not interchangeable (e.g. *pop music, soda cracker*) and by then subtracting these counts from the overall counts. Future research will explore these possibilities. Nonetheless, there are hundreds of lexical alternation variables that have never been mapped in American English that can be observed using this method. It is now possible for these variables to be measured quantitatively in a fraction of the time that it would take to conduct a traditional categorical dialect survey—in a matter of days as opposed to a matter of years.

Aside from the methodological importance of this study, it is important to consider the general results of this analysis, even though the set of variables is small. It is particularly notable how so many variants exhibit the same basic pattern, contrasting the East with the West with the approximate border between these two regions following the Ohio River and the Lower Mississippi River. This basic patterns is exhibited by 12 of the 25 variants: *running shoes, frosting, icing, cut the grass, mow the lawn, garage sale, yard sale, rummage sale, water fountain, drinking fountain, grandmother, grandma*. Furthermore, because this pattern is found in both datasets, it cannot be discounted as a result of focusing on the newspaper register. Rather, this pattern appears to represent a strong regional distinction between the language of the eastern and the western United States across registers.

This pattern is particularly notable because it is quite different from the north-south pattern that has been repeatedly identified in previous American dialect surveys, including both lexical surveys (Kurath, 1949; Carver, 1987) and phonetic surveys (Labov et al, 2006) (although see Grieve (2011) for similar patterns in contraction rate). Although there is some disagreement between previous American dialect surveys in regard to the existence and location of the Midland dialect region, in all cases the

basic finding is that the strongest pattern of regional variation in American English is a north-south divide, especially in the Eastern United States. Such patterns are visible here, especially in the map for *trash can/garbage can* alternation, but more often than not the North and the South are clustered together on the East Coast. These results are not contradictory—there is no reason to assume that all linguistic variables must pattern the same—but this finding challenges the traditional taxonomy of American dialect regions and results in a more complex picture of regional linguistic variation in American English than is commonly acknowledged.

Finally, the method introduced here also appears to be one of the most successful applications of commercial search engines for the collection of linguistic data—a practice that has recently been criticized in the literature (Kilgarriff, 2006; Lüdeling, Evert & Baroni, 2006; Baroni and Kilgarriff, 2006; Fletcher, forthcoming). Among other issues, mining Google hit counts has been criticized on the grounds that register variation cannot be controlled, that pages can be repeated and thus counted more than once, that the number of searches that can be made per day is limited, that the data is not annotated, and that search engines count pages containing particular strings rather than the strings themselves. Some of these issues have been addressed here. The use of site-restricted web searches in particular has allowed for register variation to be largely controlled. Analyzing the proportions of synonymous forms rather than analyzing the raw hit counts directly also largely neutralized the problem of counting repeated web-pages: while repeated pages will inflate the raw frequency of search strings, in general repeated pages will not effect the frequency of search strings when measured relative to other synonymous search strings. Other issues raised in these critiques have not been dealt with directly, but given the overwhelming success of the method, they are clearly not as serious as has been previously claimed. For example, the claim that search engines limit the number of searches per day is true, but it is still much quicker to search Google than to travel across a region interviewing individual informants. Similarly, although it is not possible to check the part-of-speech of strings being counted or to retrieve actual string frequencies rather than page counts, these sources of noise can be overcome through the application of advanced statistical methods, as applied here.

The fact that the using search engines such as Google to gather linguistic data has been written

off as “bad science” (Kilgarriff, 2006) is therefore not only premature but, assuming that new research has been dissuaded by these claims, counter-productive. Although it is true that commercial search engines have not always been employed successfully in previous linguistic research, this does not prove that such resources cannot or should not be exploited by linguists, as has recently been argued. Rather, as has been shown here, it is in fact both possible and productive to use commercial search engines to collect linguistic data, especially when search engines allow for linguistic data to be collected with far greater efficiency than is possible using traditional approaches. This paper has specifically described how dialect data can be gathered using search engines, but there are undoubtedly many other types of linguistic data that can be gathered using a similar approach.

Notes

. The search string must be enclosed by quotation marks to avoid searching for synonyms.

2. These measures are naturally controlled for variation in the number of webpages associated with each newspaper URL because the frequency of the forms are being measured relative to each other for just one URL at a time.

3. All maps were made in R using functions from the *maps*, *maproj* and *maptools* and *sp* packages (Bivand et al, 2008).

4. The spatial autocorrelation analysis was conducted in R using functions from the *spdep* package (Bivand et al, 2008).

5. See http://www4.uwm.edu/FLL/linguistics/dialect or http://www.tekstlab.uio.no/cambridge_survey

6. Note that the number of locations per variable for each of the HDS variables mapped here varies because locations that do not use any of the most frequent variants, which varies across variables, are excluded from each HDS map.

References

Allen, Harold B. (1973). *The Linguistic Atlas of the Upper Midwest*. Minneapolis: University of Minnesota Press.

Atwood, E. Bagby. (1962). *The Regional Vocabulary of Texas*. Austin: University of Texas Press.

Baroni, Marco and Kilgarriff, Adam. (2006). Large linguistically-processed web corpora for multiple languages. *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics*: 87–90.

Carver, Craig. (1987). *American Regional Dialects*. Ann Arbor: University of Michigan Press.

Cassidy, Fredric G. and Hall, Joan Houston. (1985). *Dictionary of American Regional English, Volume I: Introduction and A-C*. Harvard University Press.

Cassidy, Fredric G. and Hall, Joan Houston. (1991). *Dictionary of American Regional English: Volume II: D-H*. Harvard University Press.

Chambers, Jack, and Trudgill, Peter. (1998). *Dialectology*. 2nd Edition. Cambridge University Press.

Davis, Alva L. (1948). *A Word Atlas of the Great Lakes Region.* Ph.D. Dissertation. University of Michigan.

Fletcher, William H. (Forthcoming). Corpus analysis of the world wide web. To appear in Carol A. Chapelle (Ed.) *Encyclopedia of Applied Linguistics.* Hoboken, New Jersey: Wiley-Blackwell.

Grieve, Jack. (2011). A regional analysis of contraction rate in written Standard American English. *International Journal of Corpus Linguistics* 16: 514-546.

Grieve Jack, Speelman, Dirk, and Geeraerts, Dirk. (2011). A statistical method for the identification and aggregation of regional linguistic variation. *Language Variation and Change* 23: 193-221.

Hall, Joan Houston. (2002). *Dictionary of American Regional English, Volume IV: P-Sk*. Harvard University Press.

Hall, Joan Houston. (2012). *Dictionary of American Regional English, Volume V: Sl-Z.* Harvard University Press.

Hall, Joan Houston and Cassidy, Fredric G. (1996). *Dictionary of American Regional English, Volume III: I-O.* Harvard University Press.

Kilgarriff, Adam. (2006). Googlelology is bad science. *Computational Linguistics* 33: 147–151.

Kurath, Hans. (1949). *A Word Geography of the Eastern United States*. University of Michigan Press.

Kurath, Hans, Hansen, Marcus L., Bloch, Bernard, and Bloch, Julia. (1939). *Handbook of the Linguistic Geography of New England.* Providence: Brown University Press.

Labov, William, (1972). *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press.

Labov, William, Ash, Sharon, and Boberg, Charles. (2006). *Atlas of North American English: Phonetics, Phonology, and Sound Change*. New York: Mouton de Gruyter.

Lüdeling, Anke, Evert, Stefan and Baroni, Marco, (2006). *Using web data for linguistic purposes*. In Hundt, Marianne, Nesselhauf, Nadja, and Biewer, Carolin (Eds.) *Corpus Linguistics and the Web*. Amsterdam: Rodopi.

Odland, John D. (1988). *Spatial Autocorrelation*. Thousand Oaks, CA: Sage Publications.

Ord, J. K. and Getis, Arthur. (1995). Local spatial autocorrelation statistics: Distributional issues and

an application. *Geographical Analysis* 27: 286-306.

Pederson Lee, McDaniel, Susan L., and Adams, Carol M. (1986-93). *Linguistic Atlas of the Gulf States* (7 Volumes). Athens, Georgia: University of Georgia Press.

Szmrecsanyi, Benedikt. (2008). Corpus-based dialectometry: Aggregate morphosyntactic variability in British English dialects. *International Journal of Humanities and Arts Computing* 2: 279–296.

Vaux, Bert. 2003. American Dialects. In Steven Goldberg (Ed.) *Let's Go USA 2004*. Upper Saddle River, NJ: Prentice Hall.